FUNCTIONAL PLANT BREEDING JOURNAL
Scientific Journal. ISSN 2595-9433
Volume 3, Number 2, Article n. 11, July/December
D.O.I. http://dx.doi.org/10.35418/2526-4117/v2n2a1
Received: 05/28/2020 - Accepted: 06/23/2020

LINEAR, GENERALIZED, HIERARCHICAL, BAYESIAN AND RANDOM REGRESSION MIXED MODELS IN GENETICS/GENOMICS IN PLANT BREEDING

Marcos Deon Vilela de Resende[1]*, Rodrigo Silva Alves[2]

1 Embrapa Café/Universidade Federal de Viçosa, Departamento de Estatística, Campus Universitário, 36570-900, Viçosa, MG, Brasil;
2 INCT Café/Universidade Federal de Viçosa, Departamento de Biologia Geral, Campus Universitário, 36570-900, Viçosa, MG, Brasil.
** Corresponding author: Marcos Deon Vilela de Resende (marcos.deon@ufv.br).*

Abstract: This paper presents the state of the art of the statistical modelling as applied to plant breeding. Classes of inference, statistical models, estimation methods and model selection are emphasized in a practical way. Restricted Maximum Likelihood (REML), Hierarchical Maximum Likelihood (HIML) and Bayesian (BAYES) are highlighted. Distributions of data and effects, and dimension and structure of the models are considered for model selection and parameters estimation. Theory and practical examples referring to selection between models with different fixed effects factors are given using the Full Maximum Likelihood (FML). An analytical FML way of defining random or fixed effects is presented to avoid the subjective or conceptual usual definitions. Examples of the applications of the Hierarchical Maximum Likelihood/Hierarchical Generalized Best Linear Unbiased Prediction (HIML/HG-BLUP) procedure are also presented. Sample sizes for achieving high experimental quality and accuracy are indicated and simple interpretation of the estimates of key genetic parameters are given. Phenomics and genomics are approached. Maximum accuracy under the truest model is the key for achieving efficacy in plant breeding programs.

Keywords: model selection, fixed or random, comparisons with different fixed effects, HIML/HG-BLUP, FML, statistical inference.

Statistical modelling and estimation methods

Statistical and genetical modelling of experimental data plays a fundamental role in genetic improvement. Of particular importance are the precise and accurate estimation or prediction of individual genetic values and the inference about genetic control of the traits (variance components, broad and narrow sense heritabilities, repeatability, correlations and genotype × environment interaction). These guide all other activities, mainly selection and crossing. All these should be done under the truest or more correct model. High accuracy and precision provide efficiency, which together with the right model selection warrant the efficacy of the breeding program. Efficiency means to do right

(with precision and accuracy) the things, while efficacy means to do the right thing (the one that should be done, aiming to get the desired result).

At least, four classes of statistical inference (Frequentist Least Square, Restricted Maximum Likelihood, Hierarchical Maximum Likelihood and Bayesian) and five estimation methods: Least Square (LS), Restricted Maximum Likelihood/Best Linear Unbiased Prediction (REML/ BLUP), Iterative Weighted Least Squares-REML/BLUP (IWLS-REML/BLUP), Bayesian Markov Chain Monte Carlo (BMCMC) and IWLS-Hierarchical Maximum Likelihood/ Hierarchical Generalized BLUP (IWLS-HIML/ HG-BLUP), are useful in genetics and breeding, according to type and distribution of data and effects, leading to the following classes of models: Linear Fixed Model (LFM), Linear Mixed Model (LMM), Generalized Linear Mixed Model (GLMM), Bayesian Random Model (BRM) and Hierarchical Generalized Linear Mixed Model (HGLMM) (Table 1).

All models are mixed because all contain a mean and a residual variance. Searle (1971) wrote *"in point of fact, of course, all models having both mean and error terms are mixed models because the mean is a fixed effect and the errors are random"*. Then simplified denominations of the models are Linear, Generalized, Hierarchical and Bayesian.

Table 1. Data distributions, classes of models, estimation methods and statistical inference classes as applied to genetics and breeding.

Class of data	Class of model	Type of data and distribution	Estimation method*	Class of Inference
Phenotypic, genomic, phenotypic plus genomic	LFM	Balanced and continuous (Normal)	LS	Frequentist Least Square
	LMM	Continuous (Normal)	REML/BLUP	Residual Maximum Likelihood
	GLMM	Continuous (Normal), exponential family for the residual (discrete and continuous)	IWLS-REML/BLUP	Residual Maximum Likelihood
	HGLMM	Continuous (Normal), exponential family for any random factor (discrete and continuous)	IWLS-HIML/HG-BLUP	Hierarchical Maximum Likelihood
	BRM	Discrete and continuous (any)	BMCMC	Bayesian

HG-BLUP and BLUP: they are also a conditional mode (COND-MOD) estimator. * Variance/Mean parameters.

Traditionally, three classes of approaches have been used for statistical inference: frequentist (Pearsonian), likelihood (Fisherian) and Bayesian. According to these approaches, the confidence/ credibility intervals for unobservable variables can be: Fisher's fiducial (fixed interval, for fixed unknowns), frequentist (random interval, for fixed unknowns), and Bayesian (fixed interval, for random unknowns). Recently, a fourth approach, hierarchical likelihood, has emerged as a way of unifying the three classes of approaches (Lee et al., 2017).

Generally, the variables associated with traits are classified as fixed or random. Fixed variables are denoted "parameters" and no assumptions are made about their distributions. Random variables are assumed to be sampled from a probability distribution with known parameters. Estimates obtained by the Maximum Likelihood and Bayesian methods must be located in the parametric space, since if outside of that space such estimates have zero likelihood. This is not guaranteed by the Least Square estimators. For continuous variables, likelihood

is computed as the statistical density of the conditional distribution to the data sample. The statistical density *f(y)*, for a continuous variable *y*, is defined as the ordinate of the distribution function (accumulated) for a given value of *y*.

The Least Square method has limitations to handle with unbalanced data. Then, the Linear Mixed Models (LMMs) for variables with continuous Normal distribution were developed by Henderson (1952; 1973; 1975) and are implemented via BLUP and estimation of variance components by REML (so called Residual, Restricted or Reduced Maximum Likelihood) developed by Patterson and Thompson (1971) and Thompson (1973). Generalized Linear Mixed Models (GLMMs) were developed by Nelder and Wedderburn (1972) to deal with discrete variables. Lee and Nelder (1996) extended the BLUP approach to a broad class of statistical models with random effects, called Hierarchical Generalized Linear Mixed Models (HGLMMs). H stands for Hierarchical, Stratified or Structured.

In GLMMs it is assumed that the residuals may not have a Normal distribution, but the other random effects of the model follow the Normal distribution. However, this assumption is not always appropriate. An example is the situation in which the data follow the Poisson distribution and the link function specified for the residuals is Logarithmic. In this case, a more appropriate assumption for the other random factors is a Gamma distribution with a Logarithmic link function. Models in which a probability distribution and a link function can be specified for each random factor in the model belong to the HGLMMs class. Since random factors are not always hierarchically classified, an alternative name for HGLMMs is Stratified Generalized Linear Mixed Models (SGLMMs). A BLUP predictor for HGLMMs was presented by Lee and Ha (2010). For non-Normal HGLMMs, linear BLUP may not be efficient. The authors presented a combination of BLUP with Tweedie dispersion models based on Exponential distribution.

After initial works by Robertson (1955), Ronningen (1971) and Dempfle (1977), Gianola and Fernando (1986) proposed the Bayesian estimation for models of genetic evaluation. In addition to the Normal distribution adopted for the random effects (g) in the classical linear mixed model and for the likelihood of the vector of observations (y), the Bayesian approach requires assignments for the *a priori* distributions of the fixed effects and components of variance.

The attribution of non-informative or uniform *a priori* distributions for the fixed effects and components of variance is a way of characterizing an *a priori* vague knowledge about the referred effects and components. Thus, the estimation of the fixed and random effects of the Fisherian model, using the Bayesian approach, can be performed as long as non-informative prior is assigned for the fixed effects, Normal prior for the random effects and Normal likelihood for the vector of observations.

Using non-informative *a priori* distributions for the fixed effects and components of variance, the modes of the *a posteriori* marginal distributions of the components of variance correspond to the estimates obtained by REML. The paper by Gianola and Fernando (1986) was an important publication before the MCMC era. At that time the application of Bayesian methods was technically arduous and required advanced computational techniques. Beginning in 1990, statisticians introduced MCMC methods (Gelfand and Smith, 1990) and this marks the start of a new era for analysis in quantitative genetics. MCMC is especially well suited for implementing Bayesian models by sampling-based approaches to calculating marginal densities.

On the other hand, Fisherian estimation of Bayesian models can be performed via HGLMMs, with computational advantages (less time and trivial convergence criterion). HGLMMs can be fitting using their Hierarchical Likelihood (HL), which is an extension of the joint likelihood used by Henderson and consists of a joint density for observations and random effects. The estimates of fixed and random effects are derived from the maximization of HL and produce direct extensions of Henderson's mixed model equations. The components of variance are estimated by maximizing the adjusted HL profile, which is a direct extension

of REML. In this way, HGLMMs extends the familiar BLUP theory used in genetics to a broader class of models.

The class of Hierarchical Generalized Linear Mixed Models (HGLMMs) are fitted via Hierarchical Maximum Likelihood (HIML) by the Iterative Weighted Least Squares (IWLS) algorithm. This methodology allows predictions by the Hierarchical Generalized BLUP method (HG-BLUP) for random effects and estimates fixed effects by the Hierarchical Generalized Best Linear Unbiased Estimation method (HG-BLUE). The components of variance are estimated via HIML. This recently developed statistical approach for estimation, prediction, inference and model selection is very appealing.

Exploring the hierarchical nature of HL, models for the variance components of the dispersion parameters can be added one by one. A broad class of distributions can be used to model both the response variable and the random effects, a fact that increases the flexibility of the modeling. HL can also be used to derive model selection tools. The conditional Akaike Information Criterion (cAIC) is analogous to the Deviance Information Criterion (DIC) used in Bayesian statistics.

The HGLMM methodology allows also specifying y with probability distributions other than the traditional Normal, Binomial, Poisson and Negative Binomial. This can be relevant for several practical applications. For example, growth traits (diameter and height) in tree species are better described by the Weibull distribution than by Normal. Additionally, in this case, the assignment of a Gamma distribution (belonging to the family of Eulerian distributions) to y may be even more efficient, since Weibull is a particular case of the Generalized Gamma (Percontini et al., 2014).

The option of fitting the various factors of random effects under different distribution assumptions is of great interest and can be done via HGLMMs, that is, the definition of these distributions does not need to be confined only to the Normal distribution. This option can lead to greater predictive and selection efficiency, especially in plant breeding, in which the models include many factors of random effects.

The LS method does not promote the regularization (shrinkage) of the estimation process (Resende et al., 2014) and does not allow to consider the correlation between levels or effects belonging to the various factors, for example, it does not consider the correlation between levels of the effects of the treatments factor. On the other hand, REML, BAYES and HIML allow to consider these correlations.

In terms of the treatments factor, when it has a genetic connotation (comparison of individuals, for example), the correlation matrix between the levels of the factor's effects can be uncorrelated (diagonal D, which can be an identity I in the case of random effects or a null matrix in the case of fixed effects) or correlated given by three types of information: genealogical (correlation matrix A), genomic (correlation matrix G) and both simultaneously (correlation matrix H).

In terms of the animal and plant breeding, genetic selection can be carried out via: phenomic selection (genetic values predicted based on genealogy and phenotypes); genomic selection (genetic values predicted based on marker genotypes and phenotypes); geno-phenomic selection (genetic values predicted based on marker genotypes, phenotypes and genealogy, by the single step procedure via H matrix).

Additionally, a multivariate or a structured correlation matrix (longitudinal, spatial, curvilinear) can be imposed on treatment factors ($I \otimes M$, $A \otimes M$, $G \otimes M$, $H \otimes M$; where M is a matrix that describes the correlation structure) or other random effects, like residual. The combination of these four (LS, REML, BAYES and HIML) estimation methods with the four types of correlation matrix (assuming Multivariate Normal distribution of individual observations) provides the thirteen general types of statistical approaches used in genetic analyses, as shown in Table 2. Within these approaches, the following models can be fitted: Univariate, Multivariate, Longitudinal, Spatial, Curvilinear, Competitional and Survival.

Table 2. Estimation methods of variance parameters, correlation matrices and statistical models used in genetic analyses.

Estimation method of variance parameters	Correlation matrix				Model
	D	A	G	H	
LS	D-LS (ANOVA)	-	-	-	Univariate, Multivariate
REML	D-REML	A-REML	G-REML	H-REML	Univariate, Multivariate, Structured, Longitudinal, Spatial, Curvilinear, Competitional, Censored (Survival)
HIML	D-HIML	A-HIML	G-HIML	H-HIML	Univariate, Multivariate, Structured, Longitudinal, Spatial, Curvilinear, Competitional, Censored (Survival)
BAYES	D-BAYES	A-BAYES	G-BAYES	H-BAYES	Univariate, Multivariate, Structured, Longitudinal, Spatial, Curvilinear, Competitional, Censored (Survival)

Estimation and prediction of components of means via Conditional Modes (COND-MOD)

Lee and Nelder (2004) see the analysis process as consisting of two main activities: selection of the model in order to find models with good and parsimonious fitting, and predict-tion of the quantities of interest using the selected models taking into account their uncertainties. Thus, inferences about marginal responses or individual subjects belong to the prediction phase. Then, the conditional model is the basic model and any conditional model (individual prediction) leads to a specific marginal model (prediction in the mean or average).

The distinction between prediction and estimation was first reported by Lane and Nelder (1982). Prediction is a different purpose than estimation. Estimation forms the basis of predict-tions. Estimation and prediction are not the same except by chance. Lee and Nelder define predict-tion when future (unobserved) observations are "estimated" and estimation when random effects are "estimated" in data already observed. It is an estimation of unknowns in a vector *v*, which become fixed when the *y* data are observed, although possibly changing in future samples. Therefore, in this case, BLUE of the random parameters is said instead of BLUP. An unobservable future observation is not fixed given the data.

In the model for *v* in $Y_{ijk} = u + v_{ij} + \xi_{ijk}$, in which $v_{ij} = g_i + ge_{ij}$ and Y_{ijk} and ξ_{ijk} are the observation and random error, the desire to use marginal predictions ($\hat{u} + \hat{g}_i$ or $\hat{u} + \hat{g}_i + \widehat{\overline{ge}_{..}}$, for example) it is not a reason for not using conditional models. Inferences will usually be richer if conditional models are used. Care should be taken when comparing parameter estimates by different models. The conditional prediction ($\hat{u} + \hat{g}_i + \widehat{ge_{ij}}$, for example) provides confidence interval of a prediction for a potential observation given individual risk factors ($\widehat{ge_{ij}}$). This aspect is unique in conditional modeling and has wide application.

The HL uses modes and curvature to make inferences about unobservables. The term "Conditional Modes" (COND-MOD) is prefer-ble to the name BLUP because it better captures reality and makes more sense in generalized and nonlinear contexts and in Non-Linear Hierarchi-cal Models in which the values of random effects that would be estimated, would be BLUP if they were linear (that is, linear functions of observations) and unbiased. But, there is no clear attribute where they are best. For a Generalized Linear Mixed Model or a Non-Linear Mixed Model these estimates are not BLUP (Harville, 2008; Witkovský, 2012).

The BLUPs of a Generalized Linear Mixed Model provide the Conditional Modes of random effects, rather than BLUPs or Best Linear

Unbiased Predictors. They are the modes that maximize the density function of the random effects conditional in the variance-covariance parameters and in the data, that is, given the observed data and for fixed (known) values of the parameters. For the particular case of a Linear Mixed Model, these modes are also BLUPs (Resende et al., 2018).

Historically the term hierarchical was often used as a synonym for nested. However, it has been recognized within the linear model community that mixed and random effects models (nested or not) can be seen as hierarchical (in the sense of stratification). As random factors are not always hierarchically classified, an alternative term for HGLMM is Stratified or Structured Generalized Linear Mixed Model (SGLMM). These are also COND-MOD and also include Non-Linear Models.

An advantage of the mode estimator over the sample mean is that it allows the selection of the best model instead of the average model. Ma and Jorgensen (2007) advocate against the use of modal estimates for random effects and proposed the use of the Orthodox BLUP method under averages. However, Lee and Ha (2010) showed that the mode estimation of the HL function via HG-BLUP provides both better statistical precision and maintenance of the declared level of coverage probability, better than the Orthodox BLUP method.

Estimation and prediction via HIML/ HG-BLUP with simultaneous fitting of the mean and dispersion parameters

Lee and Nelder (2006) introduced the class of Double HGLMM (DHGLMM) in which random effects can be specified in both the mean and dispersion components. DHGLMMs allows modeling of the mean and variance of the variance components of random effects and residual dispersion parameters. Thus, for example, the model for residual variance includes both effects, fixed and random, on a logarithmic scale.

The distributions for the variance components are not restricted to the Inverse Chi-Square (as generally adopted in the Bayesian approach of conjugated distributions) but are also derived from the Gamma distribution with its various derived distributions. This also leads to greater flexibility in modeling. Generalized Linear and Hierarchical Generalized Linear Models allow variables distributed in the exponential family (Normal, Gamma, Poisson, Binomial) and allow a non-linear link between the observation and the linear predictor.

The REML is a special case of a Generalized Linear Model with Gamma-distributed "data" $(y - Xb)^2$ (Thompson, 2019). Extension to the multivariate case of sums of squares and cross products distributed as Wishart distribution can also be used to model data. For heritabilities, which are defined in the parametric space between 0 and 1, the best distribution is Beta, which is also defined in the space between 0 and 1 and then best describes the process.

Models with mean and variance components in the dispersion

In genetic improvement, the heterogeneity of the residual variance within families can be identified using a structural model for the variances via a Linear Log Model (Resende, 2007a). A functional form can also be used, such as variance proportional to a power function of the mean. This latter approach will be considered below using models with mean and variance components in the dispersion.

Dispersion modeling is important in statistics, for example, in the Heteroscedastic Linear Model $y \sim N[X\beta, \exp(X_{var}\beta_{var})]$, which requires REML for unbiased estimation of fixed effects on variance (β_{var}). BLUP generally refers to estimated genetic values for the average ($\hat{v}_{med}$). The model

$$y = X\beta_{med} + Zv_{med} + e,$$

traditionally assumes homoscedastic residuals, that is, $e \sim N(0, I\sigma_e^2)$. But selection can lead to an increase in residual variance. Thus, a model with residual variance heterogeneity can be recommendded (Sorensen and Waagepetersen, 2003), with residuals with the following distribution:

$$e \sim N[0, exp(X_{var}\beta_{var} + Z_{var}v_{var})],$$

or equivalently

$$e \sim N[0, exp(X_{var}\,\beta_{var})\, exp(Z_{var}\,v_{var})],$$

and genetic values are estimated for the variance ($\hat{v}_{var}$), according to model

$$y_{var} = X_{var}\beta_{var} + Z_{var}v_{var} + e_{var}.$$

Then the uniformity of the trait in the population can be increased by the selection of individuals with lower estimated genetic values for the residual variance. Thus, to obtain uniformity, the ideal is to select individuals with lower genetic values v_{var} estimated for variance and higher v_{med} estimated for the mean. The correlation $cor(v_{med}, v_{var})$ can also be estimated and, if negative, indicates that selection by v_{med} already leads to greater uniformity (lower $\hat{v}_{var}$). If positive, selection must be based on both v_{med} and $\hat{v}_{var}$, and then methods are needed to estimate both simultaneously. One example using the Gamma distribution in an experiment with Eucalyptus clones is presented below.

Heritabilities of the mean and dispersion components in the Gamma distributed data.

Variation Source	Components of variance (cv) of the mean		Components of variance (cv) of the dispersion	
	cv	e^{cv}	cv	e^{cv}
Clones	-3.332	0.036	-2.101	0.122
Residual	-2.879	0.056	-2.879	0.056
Total	-	**0.092**	-	**0.179**
h^2 (heritabilities)	-	**0.39**	-	**0.69**

It can be seen that the residual variance between clones is under genetic control with heritability of 0.69. Breeding traditionally uses only the distribution of the mean (components of means or first moments g_i) and does not use variance distribution ($\hat{v}_{var}$). But both can be used simultaneously, via ($g_{ij_{max}} = g_i + 3.09\sqrt{\hat{v}_{var_i}}$) which $g_{ij_{max}}$ indicates the maximum genotypic value of an individual *j* in the family *i* (Resende, 2015). In this case, the interest lies in greater $\hat{v}_{var}$ and a proposition of the MEAN-DISP method of selection can be done, which uses both the mean and dispersion.

This index can be used for reselection within families or populations. In this case, blocks of families are settled in new experiments for the identification of superior exceptional individual (Resende and Barbosa, 2006).

The number of plants per family *k* can be given by $n_k = (g_{ij_{max}}/g_{kl_{max}})200$, where 200 is the adequate number of individuals in the family of the best individual obtained according to the distribution of the maximum as given by Escobar et al. (2018).

Dimension and structure of models

Regarding to the dimension and structure of models, they can be classified in Univariate, Multivariate, Curvilinear, Structured and Censored, with the structures as in the Table 3.

It can be seen (Table 3) that Random Regression approaches are the full basis of genomic selection. This means that SNP prediction can be accomplished by Ridge, Bayesian and Lasso Random Regressions. This allowed the evolution of genetic predictions from family and progeny, to individual and gene (SNP) levels. According to the generic model $y = Xb + e$, where *y* is the response variable and *X* is a matrix of the marker covariates, a comparison between the LS, RR (Ridge Regression) and Lasso methods is presented below. It can be seen that the RR solution contains in addition to the LS solution the regularization parameter λ, which imposes *shrinkage* according to the process called Tikhonov regularization, which is a Gaussian process comprising n stochastic processes. If the *Lasso1* solution is negative, then the regression coefficient is declared as 0; if the *Lasso2* solution

is positive, then the regression coefficient is also declared as 0. Table 4 shows the main modelling applied to multivariate data in genetic improvement. Concerning to the types of factors and effects, the models can be classified as in the Table 5.

Table 3. Dimension in space and structure of models.

Dimension in Space	Structure
Univariate	Several factors (plots, blocks, common environment, repeatability and genotype × environment interaction), several experimental and genetic designs, several reproductive and propagation systems
Multivariate	Unstructured, AMMI, PCAMM, FAMM
Curvilinear	Random Regression, Cubic Spline, B-Spline, P-Spline, Fourier Series
Structured	Spatial, Temporal, Longitudinal, Competitional (social interaction) Repeated Measures, Autoregressive, Compound Symmetry, Exponential, Spherical, Ante-Dependence, Path Structural Equations
Censored	Survival, Longevity, Precocity
	Subtypes
Random Regression	Legendre polynomials (Random Regression); Segmented Polynomials (Cubic, B and P Splines); Ordinary Polynomials (Reaction Norms); Fractional Polynomials Autoregressive; Ridge Regression (genomics); Bayesian Regression (genomics); Lasso Regression (genomics)

AMMI: Additive Main Effects and Multiplicative Interactions; FAMM: Factor Analytical Mixed Model; PCAMM: Principal Components Analysis Mixed Model.

Comparison between the LS, RR (Ridge Regression) and Lasso methods.

Method	LS	RR	Lasso
Estimator of *b*	$\hat{b}_{LS} = (X'X)^{-1}X'Y$	$\hat{b}_{RR} = (X'X + \lambda)^{-1}X'Y$	$\hat{b}_{Lasso1} = (X'X)^{-1}X'Y - \lambda$, if the LS solution is positive $\hat{b}_{Lasso2} = (X'X)^{-1}X'Y + \lambda$, if the LS solution is negative

Table 4. Modelling applied to multivariate data.

Multivariate data	Modelling
Multivariate	Unstructured Multivariate; Principal Components; Latent Factors
Multi-Environment	Centered Principal Components (AMMI); Latent Factor Analytical Mixed Model (FAMM); Random Regression via Reaction Norms (curvilinear)
Incremental Repeated Measures	Compound Symmetry; Autoregressive; Structured Ante-Dependence (SAD)
Longitudinal Repeated Measures	Curvilinear (Random Regression via Legendre Polynomials, via Type B Segmented Polynomials (Spline), via Fourier Series)
Spatial Data	Separable Autoregressive Models AR1 x AR1; Exponential Geostatistical Model; Spherical Geostatistical Model
Time Series	Moving Averages (MA)Models; Autoregressive Models; ARMA Models; ARIMA Models; Kalman Filter (BLUP), Fourier Series

Table 5. Types of factors or effects and effects classes.

Factors or Effects	Genetical
	Residual
	Permanent
	Common Environment (plot)
	Maternal
	Contemporary Group (block)
	Macroenvironment (site)
	Genotype ×Environment Interaction
Effects classes	Fixed
	Random
	Fixed Covariables
	Polynomial Equations Covariables

Pioneering papers in Brazil on each kind of analytical model and estimation procedure

Combining all the information presented in Tables 1, 2, 3, 4 and 5, several analytical methods and estimation procedures were produced. Some pioneer papers in Brazil on each of these situations, particularly those on mixed linear modelling in plant breeding, are presented in Table 6, with the aim of showing how the practical use and the complexity of models have evolved.

Resende et al. (1990) and Resende (1991) presented some theory of the deriving of optimal selection indexes, involving various sources of information and comment on combined selection indexes involving data of individuals and their relatives. These indexes (Resende and Higa, 1994) are BLUP for the case of balanced data, as demonstrated by Resende and Fernandes (1999a). Combined selection indexes were first proposed by Lush (1945) for animals, by Wright (1962) for allogamous plants and by Weber (1982) for autogamous plants.

From 1990 to 2000 all the fundamental principles and methodology were approached, including Multivariate, Bayesian, Random Regression and Generalized Models for discrete data. Also, software's in Fortran were developed. From 2001 to 2006, due to the availability of higher computational power, it was possible to use more complex models and to fill details of the methods. From 2007 on, genomic selection came into vogue, after the advent of high density SNP (Single Nucleotide Polymorphism) genotyping. Also, HGLMMs started to gain attention.

Hypothesis tests, goodness of fit, parsimony and model selection

The hypothesis tests regarding fixed and random effects in the context of mixed models as well as the criteria for comparing models are presented in Table 7.

Hypothesis tests for fixed effects

Under REML estimation, Wald's W statistic has been recommended for testing fixed effects (Kenward and Roger, 1997). W for small samples is approximated by an F distribution. Thus, although other statistics can be introduced for the fixed effects test, Wald's statistic is attractive because it accurately reproduces the Analysis of Variance for balanced designs. If the variance components are estimated by Full Maximum Likelihood (FML), two nested models with different structure of fixed effects and with the same structure of random effects, can be compared via LRT (Gumedze and Dunne, 2011).

Hypothesis tests for random effects and criteria for model comparison

The comparison of hierarchical or nested models, with the same fixed effects structure, can be performed by Likelihood Ratio Test (LRT) or Deviance Analysis, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The comparison of non-hierarchical models, but with the same fixed effects structure, must be done using the AIC and BIC procedures.

Table 6. Some pioneering papers in Brazil on each kind of model and estimation procedure applied to genetic improvement.

Subject / References	Starting year
LMM - REML/BLUP – Univariate: Resende et al. (1993;1996;1999a,b; 2001a); Resende and Fernandes (2000); Duarte and Vencovsky (2001); Bueno Filho and Vencovsky (2000)	1993
Development of the Selegen REML/BLUP Software: Resende et al. (1994; 1999a); Resende (2007b)	1993
LMM- REML/BLUP - Unstructured Multivariate: Resende et al. (1994; 1999a,b); Resende (1999)	1994
Accuracy and experimental quality: Resende (1995); Resende and Duarte (2007)	1995
LMM - REML/BLUP - Curvilinear-Longituninal Random Regression: Resende (1997; 1999a,b); Resende et al. (2001b)	1997
BRM - Bayesian Random Models: Resende (1997; 1999a,b); Resende (2000b); Resende et al. (2001c)	1997
GLMM - IWLS-REML/BLUP for Discrete Data: Resende (2000a); Resende (2002)	2000
Use of ASReml Software: Resende (2000a)	2000
LMM - REML/BLUP - Structured - Autoregressive, Spatial: Resende and Sturion (2001; 2003); Resende (2002); Duarte and Vencovsky, 2005)	2001
LMM - REML/BLUP - Curvilinear –Splines: Resende and Sturion (2001); Resende et al. (2006)	2001
LMM - REML/BLUP - Structured - Social Competition: Resende and Thompson (2003); Resende et al. (2005)	2003
LMM - REML/BLUP - Multivariate Factor Analytical: Resende and Thompson (2003, 2004)	2003
LMM - REML/BLUP - Spatial Multivariate: Resende and Thompson (2003); Resende et al. (2006)	2003
Indices BLUP MHPRVG (Harmonic Mean of the Relative Performance of the Genetic Values) for adaptability plus stability, called Resende_indexes by Olivoto and Lúcio (2020) in a package in R: Resende (2004; 2007a)	2004
GWS - Genomic Wide Selection (RR-BLUP; G-BLUP; BAYES A, B, C and Cpi; Lasso, Blasso): Resende (2007a); Resende et al. (2008); Grattapaglia and Resende (2011); Resende et al. (2012a,b); Resende Jr. et al. (2012)	2007
HGLMM - HIML/HG-BLUP: Resende (2007a); Resende et al. (2014); Resende et al. (2018)	2007
BLUP –Autogamous (Genealogy; SIPPPG): Nunes et al. (2008); Resende et al. (2015; 2016)	2008
BLUP-Annual Allogamous: Viana et al. (2010; 2011a,b)	2010
Survival and Censored data: Resende et al. (2012), Resende et al. (2014), Santos et al. (2015; 2016)	2012
BLUP – Autogamous, Perennials, Clonal and Seminal propagation: Viana and Resende (2014)	2014

Table 7. Hypothesis tests for fixed and random effects and criteria for model comparison.

Hypothesis Tests	Effect	Asymptotic Distribution	Calculation
F	Fixed and random	F	$F = (\sigma^2_{treatment} + \sigma^2_{residual})/\sigma^2_{residual}$
WALD *n* small = F	Fixed	F	$W = \hat{\theta}^2/Var(\hat{\theta})$
WALD *n* large = LRT	Fixed	Chi-Squared	$W = \hat{\theta}^2/Var(\hat{\theta})$
Likelihood-Ratio Test (LRT)	Random	Chi-Squared	$LRT = (-2LogL)_{p+1} - (-2LogL)_p$
Akaike Information Criterion (AIC)	Random	KL	$AIC = -2LogL + 2p$
Bayesian Information Criterion (BIC)	Random	-	$BIC = -2LogL + pLog(d)$, where $d = N - r(x)$
Bayes Factor (BF)	Random	-	$BF_{12} \approx exp[-(1/2)\,\Delta_{12}]$
DIC-Bayesian	Random	-	Sampling by MCMC
C-AIC-HL	Random	-	-

$\Delta_{12} = BIC_1 - BIC_2$; KL: Kullback Leibler discrepancy; DIC-Bayesian: Bayesian Deviance Information Criterion; C-AIC-HL: conditional AIC of the Hierarchical Likelihood; $\hat{\theta}$:is the parameter;d: degrees of freedom; p: number of parameters; $r(X)$: rank of the incidence matrix X.

Likelihood-Ratio Test (LRT)

The significance of the difference in the fit of different nested models can be tested using the LRT, defined by

$$LRT = (-2LogL)_{p-1} - (-2LogL)_{p}.$$

So, it suffices to compare the difference between deviances ($-2LogL$) of the model with the *highest number of parameters - model with the lowest number of parameters*, associated with two fitted models, with the value of the probability density function (Table of χ^2) for a given number of degrees of freedom and error probability. The number of degrees of freedom is defined by the difference, between models, in the number of estimated parameters (fixed effects + components of variance). For models with the same fixed effects structure, it suffices to consider the difference in the number of components of variance.

The lower the deviance of a model, the lower the residuals of the model and the better the model. It is possible to have a negative deviance. Deviance is derived from a likelihood, which derives from Probability Density Functions (PDF). Evaluated at a certain point in the parametric space, the PDF may have a density greater than 1 due to small standard deviation or lack of variation. Likelihoods greater than 1 lead to negative deviances and are even adequate (Hall, 2014). The important thing is that the difference between deviances of two models is a positive value.

For testing close to the limit of the parametric space, Stram and Lee (1994) suggest a correction by multiplying the P value associated with χ_1^2 by 0.5, that is, suggest the use of a distribution $\chi_{0.5}^2$. In this case (mix of distributions with 1 and 0 degrees of freedom), the tabulated Chi-Square value for the 5% significance level is 2.79.

Akaike Information Criterion (AIC)

When two nested models are fitted, the one with more parameters has the highest $LogL$ and the lowest deviance ($-2LogL$). However, this is not necessarily the best model. This means that you cannot directly compare $-2LogL$ when the number of parameters varies between models. In addition to the LRT, another criterion for the selection of models is the AIC, which penalizes the likelihood by the number of fitted independent parameters.

The AIC is given by $AIC = -2LogL + 2p$, where p is the number of estimated parameters. Lower AIC values reflect a better overall fit. Thus, the AIC values are calculated for each model and the one with the lowest one (in at least 2 units, according to Cavanaugh and Neath, 2019) is chosen as the best model. There is an asymptotic equivalence between the choice of models according to the AIC criteria and cross-validation (Stone, 1977; Fang, 2011).

The AIC is related to the concepts of Kullback-Leibler information and Maximum Likelihood. Kullback-Leibler information is a physics concept for measuring the difference between the model (approximation of reality) and reality (data generation process). Akaike (1974) realized that the Log of the likelihood of a model is an estimator of the Kullback-Leibler information, however biased. And this bias is equal to the number of parameters in the model. Then, he defined the AIC as the deviance plus twice the number of model parameters. As the objective is to minimize the loss of information, the model with the smallest (in at least 2 units) AIC has the most support in the data. If the models show differences between AIC less than 2, the one with the lowest number of parameters must be selected.

The first term of the AIC can be interpreted as a model goodness of fit measure and the second term as a penalty. Thus, in the case where models with the same number of parameters are compared, it is necessary to compare only $-2LogL$ by the LRT. The advantage of AIC is that comparisons are not limited to models with a hierarchical structure of factors, a feature that makes AIC a generic tool for model selection. It can be used, for example, to compare models with errors showing different distributions.

The AIC can also be used to compare models based on different probability distributions for the trait: for example, Normal versus Gamma, Poisson versus Negative Binomial. If the models in the candidate collection are based on different distributions, then all of the terms in each likelihood must be retained when the values of AIC are compared, including constants that are not data dependent. This property of AIC is useful in applications where an appropriate distribution must be determined for the trait, in addition to the model size and structure. For this reason, AIC is ideally suited to generalized linear modeling applications (Cavanaugh and Neath, 2019). Three examples are presented below.

The DHGLM package (Lee and Noh, 2018) in R presents the components of variance (cv) on the logarithmic scale. Thus, negative values for the estimates can occur. Then, it is necessary to use the antilogarithm or exponentiation ($e^{cv} = 2.7183^{cv}$) to obtain estimates on the natural scale. The analysis of *y* as Normal led to heritability higher than that provided by the Gamma distribution. However, statistically, in terms of lower value for cAIC, the Normal model was shown to have a poorer fit ($cAIC = 1470.57$) than the Gamma model ($cAIC = 1447.39$), suggesting that the Gamma model is better. The AIC can be used to compare non-nested models with different distribution assumptions, but with the same fixed effects structure.

The analyses of *y* as Normal or Gamma led to the same heritabilities. As the cAIC was lower for the Gamma, this distribution is statistically selected. In fact, the diameter follows the Weibull distribution (special case of the Gamma distribution) (Rennolls et al., 1985) and not the Normal. It can be seen that the analysis of *y* as Binomial and Logit as link function led to greater heritability than *y* taken as Normal.

The AIC cannot be used to compare models based on different transformations of the outcome trait: for example, Log versus Square Root. Then, this criterion cannot be used to select an optimal transformation.The objective of a good criterion is to identify the fitted candidate model that is closest to the generating model in the sense of Kullback-Leibler information. AIC provides an asymptotically unbiased estimator of the expected Kullback discrepancy (Cavanaugh and Neath, 2019).

Example 1. Analysis of simulated data (*y*) using HGLMM with Normal and Gamma distributions.

Source of Variation	HGLMM-cv	e^{cv}	HGLMM-cv	e^{cv}
Genotype	3.641	38.130	-3.514	0.030
Plot	1.314	3.721	-5.405	0.004
Residual	3.019	20.471	-3.958	0.019
Total	-	62.322	-	0.053
h^2 (heritability)	-	0.61	-	0.56
Distributions	**Normal**		**Gamma**	

Example 2. Analysis of stem diameter in Acacia(*y*) using HGLMM with Normal and Gamma distributions.

Source of Variation	HGLMM-cv	e^{cv}	HGLMM-cv	e^{cv}
Genotype	-4.097	0.017	-5.129	0.006
Plot	-2.962	0.052	-4.183	0.015
Residual	-1.426	0.240	-2.212	0.109
Total	-	0.308	-	0.131
h^2 (heritability)	-	0.05	-	0.05
Distributions	**Normal**		**Gamma**	

Example 3. Analysis of survival data (0 and 1) in Acacia using HGLMM with Binomial data, Logit link function (Logistic distribution for the error or variable in latent scale) and Normal for random genotype and plot effects.

Source of Variation	HGLMM-cv	e^{cv}	HGLMM-cv	e^{cv}
Genotype	-5.498	0.004	-2.016	0.133
Plot	-4.664	0.009	-1.501	0.223
Residual	-1.761	0.172	0	1
Total	-	0.185	-	1.356
h^2 (heritability)	-	0.02	-	0.10
Distributions	**Normal**		**Binomial**	

Bayesian Information Criterion (BIC)

Another approach is BIC (Schwarz, 1978), which is given by

$$BIC = -2LogL + pLog\ (d),$$

where $d = N - r(X)$ is the number of degrees of freedom of the residual; *N* is the total number of observations and $r(X)$ is the rank of the incidence matrix (*X*) of the fixed effects. The BIC is calculated for each model and the one with the lowest value (in at least 2 units, according to Neath and Cavanaugh, 2012) is chosen as the best model. It can also be used when the models have no hierarchical structure. However, the models must have the same fixed effects structure. Logically all LRT, AIC and BIC depend on the same basic quantity $-2LogL = Deviance$.

The AIC and BIC have the same goodness-of-fit term, but the penalty terms differ on the manner in which the dimension *p* is incorporated: BIC employs a complexity penalization of $pLog(d)$ as opposed to $2p$. As a result, BIC tends to choose more parsimonious fitted models than those selected by AIC. The differences in selected models may be pronounced in large-sample scenarios (Cavanaugh and Neath, 2019).

Bayes Factor (BF)

In the Bayesian framework, the analogous to LRT, AIC, SEP (Standard Error of Prediction) and Confidence Interval, are the BF, Bayesian DIC, Standard Deviation-MCMC (SD-MCMC) and Bayesian Credible Interval (BCI), respectively. Other option to BCI is the Highest Posterior Density (HPD).

The Bayes Factor for comparing models 1 and 2 can be approximated by

$$BF_{12} \approx exp\ [-(1/2)\Delta_{12}],$$

where $\Delta_{12} = BIC_1 - BIC_2$ (Neath and Cavanaugh, 2012). The strength of evidence in terms of BF can equivalently be stated in terms of BIC. Consider a comparison between models 1 and 2, as quantified by the BIC difference Δ_{12}, being the model 2 with the smaller value of BIC. As BIC approximates a transformation of a model's posterior probability, one can perform model evaluation by transforming BIC back to a probability (Neath and Cavanaugh, 2012). Significant effects have a $BF < 0.01$, or $Log_e BF < 0$, which provide decisive evidence against the model that considered an effect equal to zero. Additionally, no significant difference is detected with $BF > 1$.

Good modelling must also take into account two relevant statistical principles: hierarchy and sparsity. According to the hierarchical principle the terms of lower order (main factors and double interactions) are generally more important than those of larger order (triple interaction, etc.). Higher order interaction generally contributes little to the explanation of a phenomenon and should not be included in the model. Sparsity refers to statistical parsimony, according to which few terms explain most of the information and the model must be kept as simple as possible.

Comparison and selection between statistical models with different factors of fixed and random effects via REML and Full ML (FML)

In the generic model

$$y = Xb + Zg + e,$$

where *y* is the vector of the data, *b* is the vector of the fixed effects, *g* is the vector of the random effects, *e* is the vector of residuals; *X* and *Z* are the incidence matrices for *b* and *g*, respectively; and $V = Var(y)$ is the variance-covariance matrix of the vector of data *y*.

The difference between the deviances of two models with different fixed effects does not provide an adequate statistical test for random effects. This is due to the fact that the Residual Likelihood (function of $y - Xb$) is maximized and not the likelihood of the original data (function of *y*, the Full Likelihood). Residual Likelihood (RL) refers to the likelihood of data after projection into the residual space and, therefore, two different models regarding fixed effects refer to two different projections and, consequently, correspond to different datasets (as if they were different variables) in which the same random factors are estimated.

In the REML method, only the portion of the likelihood that is invariant to the fixed effects (specified in vector *b*) is maximized. Thus, the components of variance are estimated without being affected by the fixed effects of the model and the degrees of freedom referring to the estimation of the fixed effects are considered, producing unbiased estimates. The REML method divides the data into two parts: contrasts of fixed effects; and error contrasts (that is, all contrasts with zero expectation) which contain information only about the components of variance. Only the contrasts of errors [full residuals $(y - Xb)$] are then used to estimate the components of variance, since they contain all available information about the variance parameters. This is done by projecting the data into the residual space or vector space of the error contrasts. The projected data has LogL given by:

$$-2RL = [N - r(X)]Log2\pi - Log|X'X| + Log|X'V^{-1}X| + Log|V| + \left(y - X\hat{b}\right)'V^{-1}\left(y - X\hat{b}\right),$$

where *N* is the number of observations and $r(X)$ is the rank of the fixed effects incidence matrix. The variance components are then estimated by maximizing the logarithm of the RL function of the projected data (Resende, 2007a; Resende et al., 2014). The LogL of the original data (Full Likelihood) is given by:

$$-2FL = NLog2\pi + Log|V| + (y - Xb)'V^{-1}(y - Xb).$$

The RL function has additional terms in relation to Full Likelihood (FL). The only additional relevant term for the estimation of variance components is $Log|XV^{-1}X|$, which effectively removes the degrees of freedom used in estimating fixed effects. This difference between RL and FL exactly reflects the difference between REML and ML (Resende, 2007a; Resende et al., 2014). Ignoring the constant terms, we have

$$RL = -(1/2)Log|XV^{-1}X| - (1/2)Log|V| - (1/2)\left(y - X\hat{b}\right)'V^{-1}\left(y - X\hat{b}\right)$$

and

$$FL = -(1/2)\,Log|V| - (1/2)(y - Xb)'V^{-1}(y - Xb).$$

For the comparison between models with different fixed effects structures, FL should be used, which can then be computed from REML by $FL^* = RL - (1/2)\,Log|(XV^{-1}X)^{-1}|$ (Verbyla, 2019). It follows the tests of random and fixed effect factors in some situations (S), where FL is Full Likelihood and RL is Residual Likelihood.

Random effect factors testing		
S1	Random effects (*g*) in nested models, with the same fixed effects and the same distribution for *y*: LRT_{RL}, AIC_{RL}, BIC_{RL}	$Model\ 1: y = Xb + Zg + e$ $Model\ 2: y = Xb + e'$ $Deviance = -2LogRL$ $LRT_{RL} = Deviance\ model\ 2 - Deviance\ model\ 1$
S2	Random effects (*g*) in non-nested models with the same fixed effects and the same distribution for *y*: AIC_{RL}, BIC_{RL}	$Model\ 1: y = Xb + Zg + e,$ $with\ relationship\ matrix\ A\ for\ example$ $Model\ 2: y = Xb + Zg' + e',$ $with\ relationship\ matrices\ G\ or\ H\ for\ example$ $Deviance = -2Log\ RL$
S3	Random effects (*g*) in nested models with different fixed effects and same distribution for *y*: AIC_{FL}, BIC_{FL}	$Model\ 1: y = Xb + Zg + e$ $Model\ 2: y = Ju + e'$ $Deviance = -2Log\ FL$
S4	Random effects (*g*) in non-nested models with different fixed effects and same distributions for *y*: AIC_{FL}, BIC_{FL}	$Model\ 1: y = Xb + Zg + e,$ $with\ relationship\ matrix\ A\ for\ example$ $Model\ 2: y = Ju + Zg' + e',$ $with\ relationship\ matrices\ G\ or\ H\ for\ example$ $Deviance = -2Log\ FL$
S5	Random effects (*g*) in nested models, with the same fixed effects and different distributions for *y*: LRT_{RL}, AIC_{RL}, BIC_{RL}	$Model\ 1: y = Xb + Zg + e$ $Model\ 2: \Upsilon = Xb + e'$ $Deviance = -2Log\ RL$ $LRT_{RL} = Deviance\ model\ 2 - Deviance\ model\ 1$
Fixed effects factor testing		
S6	Fixed effects in nested models with different fixed effects, same random effects and same distributions for y: LRTFL, AICFL, BICFL	$Model\ 1: y = Xb + Zg + e$ $Model\ 2: y = Ju + Zg + e'$ $Deviance = -2Log\ FL$ $LRT_{FL} = Dev2 - Dev1$
S7	Fixed effects in nested or not models, with different fixed effects, different random effects and same distributions for y: LRTRL, AICFL, BICFL	$Model\ 1: y = Xb + Zg + e$ $Model\ 2: y = Ju + e'$ $Deviance = -2Log\ FL$ $LRT_{FL} = Dev2 - Dev1$

Definition of fixed or random effects by analytical approach using FML

An analytical approach can be used to define fixed or random effects. An experimental example evaluating genotypes (*g*) in a complete block design, with five replicates single-tree plots is shown below. The following criteria are considered: LRT (item 1), AIC (item 2), BIC (item 3) and BF (item 4).

Three types of effects were compared for the block effects (*b*): null (CRD - Completely Randomized Design), random (CBD-R-Completely Block Design-Random) and fixed (CBD-F-Completely Block Design-Fixed). In the model for phenotypes (*y*), *u* is the general mean and *e* is the vector of random errors. The quantities *vc* and *fe* are the numbers of variance components and of fixed effects, respectively.

Results are presented for REML and FML (Full Maximum Likelihood). FML is the adequate approach to be used for comparing these models with different fixed effects of blocks. In the case 1 we conclude that the best approach (CBD-F-FML) is to consider the block effects as fixed (lower deviance of 626.8).If the REML approach

were to be considered the selected approach (CBD-R-REML) would be take the block effects as random (lower deviance of 646.29). That would lead to the wrong inference for block effects and wrong choice of the model. This emphasizes the importance of using the new FML approach. The same procedure can be used also for the genotype effects.

Using the AIC criterion (case 2) the same (block as fixed effects) conclusions can be made. The best model and effects assigned are provided by the FML approach. The lower AIC is 649.5. For the BIC criterion (case 3) a different (block as random effects) conclusion was inferred. The lower BIC was 669.0, as provided by FML approach.

For the BF criterion (case 4) we can see that, for the FML approach, all BF contrasts were significant. So, all models differ one from another. And the best model choice is as inferred by the BIC criterion, which is to take block effects as random. The BIC criterion may be the best choice as it provides a statistical formal test (BF) for the effects.

The conclusion for this example is to take block as random effects. Such a choice comes from (item 3) the fact that BIC model 2 shows lower value than that for BIC model 1 and BIC model 3, with difference between them higher than 2; and these differences were significant by the BF test (item 4).

1 DEV	Design	Model	DEV	Number of parameters
REML-DEV	CRD-REML	y=Xu+Zg+e	672.37	2 vc
	CBD-R-REML	y=Xu+Zg+Wb+e	646.29	3 vc
	CBD-F-REML	y=Xb+Zg+e	654.13	2 vc + 5fe
FML-DEV	CRD-FML	y=Xu+Zg+e	666.1	2 vc
	CBD-R-FML	y=Xu+Zg+Wb+e	641.5	3 vc
	CBD-F-FML	y=Xb+Zg+e	626.8	2 vc + 5fe

2 AIC	Design	Model	AIC	Number of parameters
REML-AIC	CRD-REML	y=Xu+Zg+e	678.4	2 vc
	CBD-R-REML	y=Xu+Zg+Wb+e	654.3	3 vc
	CBD-F-REML	y=Xb+Zg+e	668.1	2 vc + 5fe
FML-AIC	CRD-FML	y=Xu+Zg+e	672.1	2 vc
	CBD-R-FML	y=Xu+Zg+Wb+e	649.5	3 vc
	CBD-F-FML	y=Xb+Zg+e	640.8	2 vc + 5fe

3 BIC	Design	Model	BIC	N-r(X)	Number of parameters
REML-BIC	CRD-REML	y=Xu+Zg+e	693.0	975-1	2 vc
	CBD-R-REML	y=Xu+Zg+Wb+e	673.8	975-1	3 vc
	CBD-F-REML	y=Xb+Zg+e	702.3	975-5	2 vc
FML-AIC	CRD-REML	y=Xu+Zg+e	686.7	975	2 vc + 1fe
	CBD-R-REML	y=Xu+Zg+Wb+e	669.0	975	3 vc + 1fe
	CBD-F-REML	y=Xb+Zg+e	675.0	975	2 vc + 5fe

4 Bayes Factor (BF)	Design	Model	Δ	$\hat{\Delta}$	$m = -(1/2)$	$BF \approx exp(m)$
REML-BIC	CRD-REML	y=Xu+Zg+e	BIC_1-BIC_2	19.2	-9.6	7×10^{-5}
	CBD-R-REML	y=Xu+Zg+Wb+e	BIC_3-BIC_2	28.5	-14.25	6×10^{-7}
	CBD-F-REML	y=Xb+Zg+e	BIC_1-BIC_3	-9.3	4.65	104. 6
FML-BIC	CRD-FML	y=Xu+Zg+e	BIC_1-BIC_2	17.7	-8.85	1×10^{-4}
	CBD-R-FML	y=Xu+Zg+Wb+e	BIC_3-BIC_2	6.0	-3	5×10^{-2}
	CBD-F-FML	y=Xb+Zg+e	BIC_1-BIC_3	11.7	-5.85	2.9×10^{-3}

Accuracy comparison between Bayesian and Fisherian statistical models

The Bayesian accuracy could be estimated by using the same formula used in likelihood analyses, using the squared SD-MCMC instead of PEV. However, that approach is not perfect. For balanced dataset the accuracy of all individuals should give the same value. Using one example with 39 families, the Bayesian accuracy ranged from 0.43 to 0.70, with average of 0.68, mode of 0.70 and range of 0.27.

In this analysis the degrees of freedom (df) for prior distributions of variance components was 2 (which gives non-informative priors and then reproduces the REML analysis) and there was convergence (by the Geweke criterion) for all estimated genetic values and variance components, after using the MCMCglmm package (Hadfield, 2010) in R. Results are shown below.

Accuracy	REML/BLUP	MCMC$^{0.002}$	MCMC^{1}	MCMC^{2}	MCMC^{4}	MCMC^{7}
Minimum	0.67	Negative	0.33	0.43	0.49	0.54
Mean	0.67	0.66	0.67	0.68	0.68	0.68
Mode	0.67	0.69	0.70	0.70	0.69	0.69
Maximum	0.67	0.70	0.70	0.70	0.70	0.70
Mean error	0.00	0.04	0.03	0.03	0.02	0.02
Standard deviation	0.00	0.06	0.05	0.05	0.03	0.02
Range	0.00	-	0.27	0.27	0.21	0.15

0.002, 1, 2, 4 and 7: degrees of freedom.

The REML/BLUP and BLUP under MCMC estimates of variance parameters (MCMC/BLUP) analyses gave the results shown below.

Parameter	REML/ BLUP	MCMC/ BLUP$^{0.002}$	MCMC/ BLUP^{1}	MCMC/ BLUP^{2}	MCMC/ BLUP^{4}	MCMC/ BLUP^{7}
Deviance	640.62	640.73	640.67	640.67	640.67	640.67
AIC	646.62	646.73	646.67	646.67	646.67	646.67
Genetic variance	0.033	0.029	0.034	0.034	0.035	0.029
h^2	0.046	0.040	0.046	0.047	0.048	0.040
c^2	0.095	0.097	0.093	0.094	0.095	0.098
SEP	0.135	0.130	0.136	0.136	0.137	0.137
PEV	0.018	0.017	0.018	0.019	0.019	0.019
Accuracy	0.670	0.645	0.673	0.678	0.679	0.680
Bias	1.66	1.68	1.65	1.65	1.65	1.65

0.002, 2, 4 and 7: degrees of freedom.

For degrees of freedom (df) equal to 1, 2, 4 and 7 very close results were obtained for accuracy, AIC, bias and other parameters. The advantage of estimating Bayesian accuracy like this is the unique value obtained for accuracy of all individuals. MCMC/BLUP analyses can also be used for searching for better priors that could produce better results in terms of accuracy, i.e., allowing testing for informative priors.

This was tried, and the results showed that, in this case, using REML variance components as prior, no informative prior could be find other than that with 2 degrees of freedom. The results showed accuracies estimates of 0.65, 0.67, 0.68 and 0.68, for degrees of freedom 0.002, 1, 4 and 7, respectively. The last three values are very close to the 0.67 obtained with REML/BLUP.

This approach (MCMC/BLUP) is recommended for calculating accuracies in Bayesian analyses. The rationale for using this is the search for a Genuine BLUP in place of Empirical BLUP. Genuine BLUP is achieved when the parameters of the true model are known and used. In such a case, the Empirical BLUP would be replaced by the Genuine one. Genuine BLUP is unbiased, precise and has maximum accuracy. One way of seeking for that is the evaluation of the traditional BLUP machinery (Henderson mixed model equations) under parameters obtained by other approaches that can allow accessing the true model closely (Harville, 2008; Witkovský, 2012). MCMC/BLUP produces both adequate accuracy and opportunity to test of new prior distribution. It is also the only way to estimate corrected values for accuracy.

The approach of using the Bayesian standard deviation of the estimated breeding values to get accuracy, provides incorrect estimates for accuracy of every individual. And with 0.002 df even negative value of accuracy was obtained.

Another approach for computing the Bayesian accuracy was proposed by Resende et al. (2012a; 2014) and applied by Volpato et al. (2019) showing coherent and consistent results. The formula is given by $r_{\hat{g}g} = 1 - s(g)/g$, where $s(g)$ is the standard deviation of the estimated genetic value g. Other alternative is to use $r_{\hat{g}g} = 1 - s(g)/\sigma_g$, where σ_g is the squared root of the Bayesian estimate of the genetic variance, which is constant for all individuals in the population.

Genomic selection

Genomic Wide Selection (GWS) or genomic selection (GS) was proposed by Meuwissen et al. (2001) as a way to increase efficiency and accelerate the genetic improvement. GWS emphasizes the simultaneous prediction (without the use of significance tests for individual markers) of the genetic effects of thousands of genetic DNA markers (SNP) dispersed throughout the genome of an organism, in order to capture the effects of all loci (both of small and large effects) and explain all the genetic variation of a quantitative trait. Meuwissen et al. (2001) developed the SNP-BLUP procedure using the method RR-BLUP, BayesA and BayesB. The Ridge-Regression (RR) was already used by Whittaker et al. (2000) for marker selection. Haley and Visscher (1998) had already suggested the name genomic selection for selection in a whole genome scale.

An ideal method for SNP effects estimation in GWS should include three attributes: accommodate the genetic architecture of the trait in terms of genes of small and large effects and their distributions; regularize the estimation process in the presence of multicollinearity and a larger number of markers than individuals, using shrinkage estimators; perform the selection of covariables (markers) that affect the trait under analysis.

The main problem with GWS is the estimation of a large number of effects from a limited number of observations and also the collinearities arising from the linkage disequilibrium between the markers. Shrinkage estimators deal with this appropriately, treating the effects of markers as random variables and estimating them simultaneously (Resende, 2007a; Resende, 2008; Azevedo et al., 2015).

If the effects of markers are taken as fixed, it is not possible to consider the covariance between effects of markers. With a high density of markers, more than one marker will be in linkage disequilibrium with a segregating QTL. This will result in covariance between marker effects. Most markers will have no effect on a trait. Thus, the estimated effects of these empty markers will be false. This problem is greater in the case that the markers are considered to have fixed effects, because in that case, these pseudo effects will not be shrunk towards zero.

The traditional Quantitative Genetics rely on random mating population. Nowadays, with the availability of SNP markers, random mating does not need to be assumed, because breeders can track the transmission of chromosomal segments. Another assumption is linkage equilibrium in the breeding population. Once linkage among markers is accounted for in the G relationship matrix in RR-BLUP, this circumvent the need to assume linkage equilibrium.

The main methods for GWS are based on Random Regression and can be divided into three major classes: explicit, implicit and dimensionally reduced regression. In the first class, the methods RR-BLUP, Lasso, BayesA and BayesB, among others, stand out. In the class of implicit regression, the method RKHS (Reproducing Kernel Hilbert Spaces) is most popular and is a semi-parametric method. Among the regression methods with dimensional reduction, stand out the Independent Components, Partial Least Squares and Principal Components. Two new non-parametric approaches for GWS were proposed by Resende (2015) and Lima et al. (2019a,b). These methods are called triple categorical regression (TCR) and Delta-p and proved to be efficient.

The explicit regression methods are divided into two groups: (i) penalized estimation methods (RR-BLUP, Lasso); (ii) Bayesian estimation methods (BayesA, BayesB, fast BayesB, BayesCπ, BayesDπ, Bayesian regression, BayesR, BayesRS, Blasso, IBlasso and others). The best and most effective in practice are RR-BLUP (via G-BLUP single step) and BayesB (Mrode et al., 2010; Mrode, 2014; Visscher et al., 2006; 2008; 2010). Each method without covariate selection has its similar with covariate selection. Thus, there are the pairs without and with selection: BayesA - BayesB; BRR - BayesCπ; Blasso - IBlasso.

The RR-BLUP is an equivalent model to G-BLUP, which is the BLUP method at individual level with the genealogical relationship matrix A changed to a genomic relationship matrix G. The equivalence between these two methods was given by Habier et al. (2007) and also by Van Raden (2008). The G-BLUP and RR-BLUP are equivalent when the number of QTL is large and no major QTL is present. The use of the matrix G based on markers was already used by Bernardo (1994), Nejati-Javaremi et al. (1997) and Fernando (1998).

A single-step BLUP using simultaneously phenotypic, genotypic and genealogical information, called H-BLUP single-step, was proposed by Misztal et al. (2009), using an H matrix composed by the A and G matrices. The idea of the H-BLUP was already given by Fernando (1998).

The cut-off point for including a marker in the analysis can be given by $MAF = (1/2N)^{1/2}$; this comes from the standard deviation of a proportion, given by$(pq)^{1/2}/(2N)^{1/2}$, where N is the number of genotyped individuals, meaning that the lower N the greater needed to be the MAF for accurate estimation of the marker effect (Resende, 2015).

A refinement of genomic selection can be achieved by using QTNs instead of SNPs. The evolution of genomic technology is predictable and the causal mutation of a genetic variation at the nucleotide level (QTN) could be accessed in the near future. Thus, genomic selection can be improved by the direct use of QTNs instead of SNPs.

The use of QTNs will bring the following advantages (Weller, 2016): GWS will not depend on the linkage disequilibrium as the QTN will be accessed directly and not via markers, this will increase the durability of the genomic prediction, which will also be useful in the long run; the genomic prediction may have validity (transferability) across different populations and species in the same genus; the genomic prediction will use specific QTNs for each trait, unlike G-BLUP via SNPs, which uses the same G relationship matrix for all traits; the multiple-trait selection indexes will directly weight the QTNs and not the phenotypic traits; genomic selection may use a smaller number of generations (only the last ones) for the composition of G matrix, this will bring greater genetic gain and less mass of data to be processed; the allele frequencies of the QTNs will be accessed directly and not via linkage disequilibrium with SNPs.

Analytical statistics

In general, a complete statistical analysis encompasses the following activities: the model selection; the estimation/prediction of components of means (genotypic values); the estimation of components of variance (genotypic variability); the application of hypothesis tests; the inferences on accuracy (square root of the reliability of the selection); the inferences on bias; and the inferences on estimation/prediction precision (Resende, 2007b).

For example, in the context of mixed models, performing these activeties involves BLUP predicttion, REML estimation, deviance analysis, computation of prediction accuracy and prediction error variance, respectively. In the REML/BLUP procedure, the bias is assumed to be null, as these estimators/predictors belong to the class of the Best Linear Unbiased Estimators/Predictors (BLUE/ BLUP). In the scientific articles, the results from these activities should be interpreted and discussed. In the genetics area the following route can be followed.

Model selection	The best biological/statistical model should be selected by comparing several candidate models, based on information criteria, disparity measures or statistical distances between probability distributions or models (contrasts between data generating and candidate models). This activity is the first one and is essential, meaning that the best individual cannot be selected from the wrong model. Commonly used information criteria are the Kullback-Liebler (KL), Akaike (AIC) and Bayesian (BIC, which is related to the Bayes Factor).
Hypothesis tests	Inferences on significance of the genetic variability (σ^2_g), by the deviance analysis (LRT) or F test from Analysis of Variance in the balanced case.
Variance components	Their proportions allow inferences on genetic control, heritabilities, repeatability, ge interaction, correlations between traits and coefficients of variation.
Components of means	They provide information on genetic values and genetic gain with selection.
Precision of the prediction	PEV (prediction error variance), from which we can calculate the relation PEV/σ^2_g (with parameter space between 0 and 1) and also the value of $F = \sigma^2_g/PEV$. This comes from the squared accuracy estimator $r^2_{\hat{g}g} = 1 - (PEV/\sigma^2_g) = 1 - (1/F)$. For a fixed effects model, $r^2_{\hat{g}g} = 1 - (1/F)$ has a connotation of an adjusted determination coefficient, which is similar to a broad sense heritability at genotype mean level (reliability).
Accuracy of the prediction	Correlation between predicted and true genotypic values, with parameter space between 0 and 1.
Bias of the prediction $\hat{y}$	given as a function of the regression $[\beta(y, \hat{y})]$ of data y on $\hat{y}$, where $\beta = 1$ is the ideal and indicates no contribution of the angular coefficient β to bias. In this sense, comparisons of the models should be based on the modulus of $[1 - \beta(y, \hat{y})]$.

Genotype selection

The genetic selection should be based on BLUE, BLUP, HG-BLUE, HG-BLUP or COND-MOD in the context of the random and mixed effects models. For the selection of genotypes in the context of fixed effects models, multiple comparison should be done by the Newman-Keuls test and not by Tukey. The Newman-Keuls test has much higher power and type I error rates similar to the Tukey test. Thus, the t-test, Duncan and Tukey, widely used in Brazil, are not the most recommended and should only be used with cautions. The Newman-Keuls test, little used in Brazil, is highly recommended in view of the favorable rates of type I error, the relative high power and the intermediate rigor. Thus, it can be used without much care. In reality, this test has been widely used (in detriment of the others) by the French researchers and also in perennial plant improvement in African countries. French literature adopts the Newman-Keuls test as a standard in place of Tukey test. Example of calculus is presented in Resende (2002; 2007a).

Apart from this, the statistical machinery for doing all the analyses is the mixed model methodology by REML/BLUP, HIML/HG-BLUP and Bayesian estimation (Blasco, 2001; Sorensen and Gianola, 2002; Resende et al., 2014; 2018). The Selegen REML/BLUP Software (Resende, 2016), ASReml Software (Gilmour et al., 2015), Echidna Mixed Model Software (Gilmour, 2019) and some R packages (R Development Core Team, 2018) can be used.

In genetics, studies on diversity (genetic relationship coefficients, inbreeding coefficients, effective population size (N_e), entropy, genetic distances and multivariate clusters) complement the inferences (Resende, 2015).

Sample size and accuracy in plant breeding

Experimental quality and selective accuracy

The quality of the genotypic evaluation should preferably be inferred based on accuracy ($r_{\hat{g}g}$). In balanced experiments, Snedecor's F statistics can also be used, as shown in the table presented by Resende and Duarte (2007). Being $r_{\hat{g}g} = (1 - 1/F)^{1/2}$, the mathematical expression that relates the appropriate values of F to the required accuracy, is given by: $F = 1/\left(1 - \hat{r}_{\hat{g}g}^2\right)$. To achieve an accuracy of 90%, an F value equal to 5.26 must be pursued. Thus, this should be a reference value in experiments for evaluating VCU tests. This value is independent of the species and trait evaluated and can be considered as a standard value for any species. This statistic contemplates, simultaneously, the coefficient of experimental variation (CV_e), the number of replications (n) and the coefficient of genotypic variation (CV_g). The expression $F = 1 + \left(nCV_g^2/CV_e^2\right)$ shows this. Although traditionally used to evaluate experimental quality, the coefficient of experimental variation alone is not adequate for this. The three parameters are necessary, because the accuracy depends on them simultaneously, as can be seen by the alternative expression

$$\hat{r}_{\hat{g}g} = \left\{1/\left[1 + \left(CV_e^2/CV_g^2\right)/n\right]\right\}^{1/2}.$$

For the selection process in breeding programs, accuracy values above 70% should be pursued. This is equivalent to F values approximately greater than 2. Therefore, F values less than 2 provide low selective accuracy.

Another statistic commonly calculated in the context of genotypic evaluation, as proposed by Vencovsky (1987), is the coefficient of relative variation ($CV_r = CV_g/CV_e$). By fixing the number of replications or individuals per treatment, the CV_r magnitude can be used to infer about the accuracy and precision in the genotypic evaluation. With $n = 2$, a $CV_r > 1$ provides high accuracy.

In terms of individual (perennials) or plot (annuals) h^2, F is given by $F = 1 + nh^2/(1 - h^2)$, and $F = 5.2632$ is achieved, for example, with $n = 6.39$, for $h^2 = 0.4$. It can be inferred that with $h^2 = 0.4$, and $n = 6$ provides high accuracy.

Required sample sizes for treatments effects detection

High reliability and accuracy can be achieved by using adequate number of replications or individuals (n) per treatment and of repeated measures (m). This should be determined according to the heritability (h^2) and repeatability (ρ) of the traits. The quantities n and m can be given by

$$n = r_{\hat{g}g}^2(1 - h^2)/\left[h^2(1 - r_{\hat{g}g}^2)\right]$$

and

$$m = r_{\hat{f}f}^2(1 - \rho)/\left[\rho(1 - r_{\hat{f}f}^2)\right],$$

where $r_{\hat{g}g}^2$ and $r_{\hat{f}f}^2$ are the reliabilities (squared accuracy) of genetic and phenotypic values, respectively. For a trait with $h^2 = 0.20$ and $\rho = 0.40$, n should be 4 and 17, and m should be 2 and 7, for a targeted accuracy of 70% and 90%, respectively. The number of replications can also be given by $n = (F - 1)(1 - h^2)/h^2$, where F is 5.26 for a desired accuracy value of $r_{\hat{g}g} = 0.90$ (Resende and Duarte, 2007).

Statistical books provide the general expression to calculate the required sample size (n) which is $n = \left[\left(z_\alpha + z_\beta\right)^2\sigma_D^2\right]/\delta^2$, where z_α and z_β are values of the accumulated distributions function of type I (α) and type II (β) errors, under unilateral hypothesis tests; σ_D^2 is the

variance of the difference between two treatments means; and δ is the size of the real difference between two means which are intended to be declared as significant.

The quantity $(1-\beta)$ is the probability (power) that the experiment shows a significant difference between treatments means. Powers of 80% and 90% are common and adequate in practice. The variance of σ_D^2 is function of the residual variance (given as a function of $1-h^2$) and δ^2can be taken as the squared contrast between one effect and the zero point of mass (given as a function of h^2). Wearden (1959) used something similar to this. Comparing $n = \left(z_\alpha + z_\beta\right)^2(1-h^2)/h^2$ with $n = (F-1)(1-h^2)/h^2$ given before, we have $(F-1) = \left(z_\alpha + z_\beta\right)^2 = NCP$, which is the non-centrality parameter. Values of $\left(z_\alpha + z_\beta\right)^2$ were given by Snedecor and Cochran (1967) as below:

$(1-\beta)$	Unilateral tests $\left(z_\alpha + z_\beta\right)^2$ significance level α		
	0.01	**0.05**	**0.1**
0.80	10.0	6.2	4.5
0.90	13.0	8.6	6.6
0.95	15.8	10.8	8.6

With $\alpha = 5\%$ and $\beta = 90\%$, $NCP = 8.6$ and $F = 9.6$. So, $r_{\hat{g}g}^2 = 0.90$ and $r_{\hat{g}g} = 0.95$; with $\alpha = 5\%$ and $\beta = 80\%$, $NCP = 6.2$ and $F = 7.2$. So, $r_{\hat{g}g}^2 = 0.86$ and $r_{\hat{g}g} = 0.93$; and with $\alpha = 5\%$ and $\beta = 80\%$, $NCP = 4.5$ and $F = 5.5$. So, $r_{\hat{g}g}^2 = 0.82$ and $r_{\hat{g}g} = 0.91$. In this way, an accuracy of 90% is associated with α equal to 10% and β equal to 80%, among other combinations of α and β. A summary of these results is presented in Table 8.

Table 8. Significance level and power of t test associated with required accuracy levels of 0.90, 0.93 and 0.95.

Accuracy $(r_{\hat{g}g})$	$r_{\hat{g}g}^2$	Significance (Type I Error: α)	Confidence (1-α)	Power (1-β)	Type II Error (β)	F test
0.91	0.82	0.10	0.90	0.80	0.20	5.5
0.93	0.86	0.05	0.95	0.80	0.20	7.2
0.95	0.90	0.05	0.95	0.90	0.10	9.6

It can be seen that to perform an experiment with desired power of the F-test of 0.90 and significance of 0.05 we should seek for an accuracy of 0.95. In this case, the probability of detecting a true difference among genotypes is 0.90, when the significance level is set at 0.05. There is a closeness between accuracy and confidence level, as expected. Also, a relation between power and coefficient of determination ($r_{\hat{g}g}^2$) seems to hold, for such high accuracy values. The coefficient of determination is also called proportional reduction of error and is more a measure of coincidence proportion, hits or rightness (Linder, 1951).

Sample size for genomic selection

Genomic data are especially useful for genomic selection (GS), which allow selecting at plantlet stage aiming genetic gain in the adult stage (Resende et al., 2008; Grattapaglia and Resende, 2011). With GS:

$$r_{\hat{g}g}^2 = nh^2/(nh^2 + nQTL) = nh^2/(nh^2 + M_e) = nh^2/(nh^2 + 2N_eL) = nh^2/(nh^2 + L/F),$$

where n is the number of genotyped and phenotyped individuals, L is the genome size (in Morgans) species, M_e is the effective number of chromosome segments, N_e is the population efective number and F is the inbreeding coefficient of the population. For a desired $r_{\hat{g}g}^2$, h^2 and $nQTL$, n can be determined.

Classification and interpretations

Classification of the accuracy magnitudes is given in Table 9. Accuracy is linked to trait heritability. A classification of additive heritability and accuracy in terms of magnitude and their associations is presented in Table 10 (Resende, 1997).

Table 9. Adequate values of the Snedecor F statistics, for the genetic effects (cultivars), aiming to achieve a certain accuracy, and the categories of required precision in the genotypic evaluation.

Accuracy	Accuracy categories	F value	Accuracy	Accuracy categories	F value
0.99	Very high	50.2513	0.65	Moderate	1.7316
0.975	Very high	20.2532	0.60	Moderate	1.5625
0.95	Very high	10.2564	0.55	Moderate	1.4337
0.90	Very high	5.2632	0.50	Moderate	1.3333
0.85	High	3.6036	0.40	Low	1.1905
0.80	High	2.7778	0.30	Low	1.0989
0.75	High	2.2857	0.20	Low	1.0417
0.70	High	1.9606	0.10	Low	1.0101

Source: (Resende and Duarte, 2007).

Table 10. Individual additive heritability (h_a^2), accuracy for individual selection ($r_{\hat{a}a}$), maximum possible accuracy for BLUP using also the family mean ($r_{\hat{a}a\,max}$), classification of magnitudes of individual additive heritability (h_a^2 classification) and classification of accuracy magnitudes for selection of individuals ($r_{\hat{a}a}$ classification).

h_a^2	$r_{\hat{a}a}$	$r_{\hat{a}a\,max}$	h_a^2 classification	$r_{\hat{a}a}$ classification
0.01	0.10	0.51	Low $0.01 \leq h_a^2 \leq 0.15$	Low $0.10 \leq r_{\hat{a}a} \leq 0.40$
0.10	0.32	0.55		
0.15	0.39	0.58		
0.20	0.45	0.61	Moderate $0.15 < h_a^2 < 0.50$	Moderate $0.40 < r_{\hat{a}a} < 0.70$
0.30	0.55	0.66		
0.40	0.63	0.71		
0.50	0.71	0.76	High $0.50 \leq h_a^2 < 0.80$	High $0.70 \leq r_{\hat{a}a} < 0.90$
0.60	0.77	0.80		
0.70	0.84	0.85		
0.80	0.89	0.90	Very high $h_a^2 \geq 0.80$	Very high $r_{\hat{a}a} \geq 0.90$
0.90	0.95	0.95		

Source: (Resende, 1997).

It is verified that, with $h_a^2 > 0.50$, there is practically no advantage in the use of family information and the selection based only on individual information already provides a high accuracy ($r_{\hat{a}a} > 0.70$). Even for traits with low additive heritability, the use of information from relatives (more information) allows to increase the selective accuracy of the class from low to moderate. This fact highlights the importance of working with more elaborate selection methods.

Classification of magnitudes of repeatability estimates

In general, the classification of the repeatability coefficients in terms of magnitude can be performed by comparing the permanent phenotypic gain to be obtained considering one measurement (G_1) with that to be obtained assuming m measurements (G_m), by the ratio $G_1/G_m = \{[1 + (m-1)\rho]/m\}^{1/2}$.

Considering the $m = 2$ for G_m, the classifications for repeatability are as follows: high repeatability: $G_1/G_m \geq 0.90 \rightarrow \rho \geq 0.60$; medium repeatability: $0.80 < G_1/G_m < 0.90 \rightarrow 0.30 < \rho < 0.60$;and low repeatability: $G_1/G_m \leq 0.80 \rightarrow \rho \leq 0.30$.

Classification of the magnitudes of the genetic correlation coefficients

A classification of the magnitudes of the genetic correlation coefficients can be obtained by taking thirds of the values of the parametric

space that extends from -1 to 1. Thus, we have the following classification:

Positive scale values	Negative scale values	Classification
0.0 to 0.33	0.0 to 0.33	Low
0.34 to 0.66	0.34 to 0.66	Medium
0.67 to 1.0	0.67 to 1.0	High

Correlations must be interpreted not only based on their significance, but mainly based on their magnitudes. The classes shown above are valid for genetic and phenotypic correlations between traits and also for correlation of the same variable across environments. In the latter case, the low, medium and high classes for correlations should also be interpreted as genotype × environment interaction high, medium and low, respectively. The low-class correlation denotes high interaction and, in addition, that the interaction is of the complex type (arising from the lack of correlation between genotypes across environments).

In the bivariate context, the correlation coefficient between orders, or Spearman correlation between two variables, is not strongly influenced by extreme pairs. Thus, it is robust in relation to Pearson's linear correlation coefficient. A large difference in magnitude between these two types of correlation coefficient can reveal the presence of extreme pairs of variables. However, a high Spearman correlation does not necessarily indicate that the relationship between two variables is linear. Spearman's correlation between two variables, markedly higher than Pearson's correlation, may indicate a non-linear relationship between these variables. As an example, two variables *X* and *Y*, where *Y* is given by $Y = X^2$, will present a Pearson correlation value close to 0, but a Spearman correlation value equal to 1.

Coefficient of variation

In experimental statistics, the genetic variability inherent to the experiment can be measured by the coefficient of genetic variation (CV_g), which informs about the possibility of improvement and the evolution of the trait in the population. This measure is scaled and, therefore, comparable between variables. The coefficient of phenotypic variation, when greater than 100% indicate the presence of outliers.

Weights in selection index

For constructing selection index, phenotypic traits should be heritable ($h^2 > 0.10$), adequately scaled and scored and correlated with the breeding objective. Traits can be combined in super-variables or in selection indexes describing the breeding objective. An efficient alternative for calculating the economic weights w_i refers to the use of genetic correlations between each trait *i* and the objective trait *j* of the improvement ($r_{g_{ij}}$). In this case, w_i is given by

$$w_i = r_{g_{ij}} / \sum_{i=1}^{n} r_{g_{ij}},$$

that is, it is equivalent to the correlation as a proportion of the sum of the correlations involving the *n* variables and the objective trait.

Genotype x environment interaction and genotype correlation across environments

The genotype correlation across environments (r_{ge}) can be expressed alternatively according to the proportion $P = \sigma^2_{ge}/\sigma^2_g$, by means of

$$r_{ge} = \sigma^2_g/(\sigma^2_g + P\sigma^2_g) = 1/(1+P).$$

With $P = 0.5$, we have $r_{ge} = 0.67$, which is a high value of genetic correlation.

Thus, it can be inferred that when the ratio of the variance of the interaction/genetic variance free from interaction is less than 0.5, the interaction is not problematic for the breeder, as it will lead to a high correlation. When $P > 0.5$, the interaction can be problematic for the breeder, implying losses of gain with indirect selection (selection in one place aiming at gain in another). There is also the equality

$$P = \sigma^2_{ge}/\sigma^2_g = (1 - r_{ge})/r_{ge},$$

where $(1 - r_{ge})$ is the lack of correlation.

References

AKAIKE, H. 1974. A new look at the statistical model identification. **IEEE Transaction on Automatic Control**, 19:716-723.

AZEVEDO, C.F.; RESENDE, M.D.V.; SILVA, F.F.; VIANA, J.M.S.; VALENTE, M.S.F.; RESENDE JR., M.F.R.; MUÑOZ, P. 2015. Ridge, Lasso and Bayesian additive dominance genomic models. **BMC Genetics**, 16:1-13.

BERNARDO, R. 1994. Prediction of maize single-cross performance using RFLPs and information from related hybrids. **Crop Science**, 34:20-25.

BLASCO, A. 2001. The Bayesian controversy in animal breeding. **Journal of Animal Science**, 79:2023-2046.

BUENO FILHO, J.S.S.; VENCOVSKY, R. 2000. Alternativas de análise de ensaios em látice no melhoramento vegetal. **Pesquisa Agropecuária Brasileira**, 35:259-296.

CAVANAUGH, J.E.; NEATH, A.A. 2019. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. **Computational Statistics**, 11:2-11.

DEMPFLE, L. 1977. Relation entre BLUP (Best Linear Unbiased Prediction) et estimateurs bayesiens. **Annales de Génétique et Sélection Animale**, 9:27-32.

DUARTE, J.B.; VENCOVSKY, R. 2001. Estimação e predição por modelo linear misto com ênfase na ordenação de médias de tratamentos genéticos. **Scientia Agricola,** 58:109-117.

DUARTE, J.B.; VENCOVSKY, R. 2005. Spatial statistical analysis and selection of genotypes in plant breeding. **Pesquisa Agropecuária Brasileira**, 40:107-114.

ESCOBAR, J.A.D.; RESENDE, M.D.V.; AZEVEDO, C.F.; SILVA, F.F.; 2005, M.H.P.; NUNES, A.C.P.; ALVES, R.S.; NASCIMENTO, M. 2018. Teoria de valores extremos e tamanho amostral para o melhoramento genético do quantil máximo em plantas. **Revista Brasileira de Biometria**, 36:108.

FANG, Y. 2011. Asymptotic equivalence between cross-validations and Akaike information criteria in mixed-effects models. **Journal of Data Science**, 9:15-21.

FERNANDO, R.L. 1998. Genetic evaluation and selection using genotypic, phenotypic and pedigree information. **Proceedings of the 6th World Congress on Genetics Applied to Livestock Production**, Armidale, Australia, p.329-336.

GELFAND, A.E.; SMITH, A.F.M. 1990. Sampling-based approaches to calculating marginal densities. **Journal of the American Statistical Association**, 85:398-409.

GIANOLA, D.; FERNANDO, R.L. 1986. Bayesian methods in animal breeding theory. **Journal of Animal Science**, 63:217-244.

GILMOUR, A.R. 2019. **Echidna Mixed Model Software**. Orange, Australia.

GILMOUR, A.R.; GOGEL, B.J.; CULLIS, B.R.; WELHAM, S.J.; THOMPSON, R. 2015. **ASReml User Guide Release 4.1.** VSN International.

GRATTAPAGLIA, D.; RESENDE, M.D.V. 2011. Genomic selection in forest tree breeding. **Tree Genetics & Genomes,** 7:241-255.

GUMEDZE, N.; DUNNET, T. 2011. Parameter estimation and inference in the linear mixed model. **Linear Algebra and its Applications**, 435: 1920-1944.

HABIER, D.; FERNANDO, R.L.; DEKKERS, J.C.M. 2007. The impact of genetic relationship on genome-assisted breeding values. **Genetics**, 117:2389-2397.

HADFIELD, J.D. 2010. MCMC methods for multi-response Generalized Linear Mixed Models: The MCMCglmm R Package. **Journal of Statistical Software**, 33, 1-22.

HALEY, C.S.; VISSCHER, P.M. 1998. Strategies to utilize marker-quantitative trait loci associations. **Journal of Dairy Science**, 81: 85-97.

HALL, B. 2014. **Bayesian Inference**. Available at: http://m-clark.github.io/docs/ld_mcmc/BayesianInference.pdf.

HARVILLE, D. 2008. Accounting for the estimation of variances and covariances in prediction under a general linear model: an overview. **Tatra Mountains Mathematical Publications**, 39:1-15.

HENDERSON, C.R. 1952. Specific and general combining ability. In: GOWEN, J. W. **Heterosis**. New York: Hafner Publishing. p.352.

HENDERSON, C.R. 1973. Sire evaluation and genetic trends. In: **Animal Breeding and Genetics Symposium in Honor of J. LUSH**. Champaign: American Society of Animal Science. p.10-41.

HENDERSON, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, 31:423-447.

KENWARD, M.G.; ROGER, J.H. 1997. Small sample inference for fixed effects from restricted maximum likelihood. **Biometrics**, 53: 983-997.

LANE, P.W.; NELDER, J.A. 1982. Analysis of covariance and standardization as instances of prediction. **Biometrics**, 38:613-621.

LEE, Y.; HA, I.D. 2010. Orthodox BLUP versus h-likelihood methods for inferences about random effects in Tweedie mixed models. **Statistics and Computing**, 20:295-303.

LEE, Y.; NELDER, J.A. 1996. Hierarchical Generalized Linear Models. **Journal of the Royal Statistical Society**, 58:619-678.

LEE, Y.; NELDER, J.A. 2004. Conditional and Marginal Models: Another View. **Statistical Science**, 19:219-238.

LEE, Y.; NELDER, J.A. 2006. Double Hierarchical Generalized Linear Models. **Journal of the Royal Statistical Society**, 55:1-29.

LEE, Y.; NELDER, J.A.; PAWITAN, Y. 2017. **Generalized Linear Models with Random Effects**. Chapman and Hall/CRC Press, Boca Raton.

LEE, Y.; NOH, M. 2018. **dhglm: Double Hierarchical Generalized Linear Models**. R package version 2.0.

LIMA, L.P.; AZEVEDO, C.F.; RESENDE, M.D.V.; SILVA, F.F.; SUELA, M.M.; NASCIMENTO, M.; VIANA, J.M.S. 2019. New insights into genomic selection through population-based non-parametric prediction methods. **Scientia Agricola**, 76:290-298.

LIMA, L.P.; AZEVEDO, C.F; RESENDE, M.D.V.; SILVA, F.F.; VIANA, J.M.S.; OLIVEIRA, E.J. 2019. Triple categorical regression for genomic selection: application to cassava breeding. **Scientia Agricola**, 76:368-375.

LINDER, A. 1951. **Statistische Methoden für Naturwissenschaftler, Mediziner und Ingenieure**. Verlag Birkhäuser, Basel.

LUSH, J.L. 1945. **Animal breeding plans**. Iowa State University Press, Ames.

MA, R.; JORGENSEN, B. 2007. Nested Generalized Linear Mixed Models: Orthodox Best Linear Unbiased Predictor Approach. **Journal of the Royal Statistical Society**, 69:625-641.

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, 157:1819-1829.

MISZTAL, I.; LEGARRA, A.; AGUILAR I. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. **Journal of Dairy Science**, 92:4648-4655.

MRODE, R.; COFFEY, M.; BERRY, D.P. 2010. Understanding genomic evaluations from various evaluation methods and GMACE. **Interbull Bulletin**, 42:52-55.

MRODE, R.A. 2014. **Linear models for the prediction of animal breeding values**. CAB International, Wallingford.

NEATH, A.A.; CAVANAUGH, J.E. 2012. The Bayesian information criterion: background, derivation, and applications. **Computational Statistics**, 4:199-203.

NEJATI-JAVAREMI, A.; SMITH, C.; GIBSON, J.P. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. **Journal of Animal Science**, 75:1738-1745.

NELDER, J.A.; WEDDERBURN, R.W.M. 1972. Generalized Linear Models. **Journal of the Royal Statistical Society**, 135:370-384.

NUNES, J.A.R.; MORETO, A.L.; RAMALHO, M.A.P. 2008. Using genealogy to improve selection efficiency of pedigree method. **Scientia Agricola**, 65:25-30.

OLIVOTO, T.; LÚCIO, A.D. 2020. metan: An R package for multi-environment trial analysis. **Methods Ecology and Evolution**, 00:1-7. https://doi.org/10.1111/2041-210X.13384.

PATTERSON, H.D.; THOMPSON, R. 1971. Recovery of inter-block information when block sizes are unequal. **Biometrika**, 58:545-554.

PERCONTINI, A.; SILVA, F.S.G.; RAMOS, M.W.A.; VENANCIO, R.; CORDEIRO, G.M. 2014. A distribuição Gama Weibull Poisson aplicada a dados de sobrevivência. **TEMA**, 15:165-176.

R DEVELOPMENT CORE TEAM. 2020. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna.

RENNOLLS, K.; GEARY, D.N.; ROLLINSON, T.J.D. 1985. Characterizing Diameter Distributions by the use of the Weibull Distribution. **Forestry**, 58:57-66.

RESENDE JR., M.F.R.; MUÑOZ, P.; RESENDE, M.D.V.; GARRICK, D.J.; FERNANDO, R.L.; DAVIS, J.M.; JOKELA, E. J.; MARTIN, T. A.; PETER, G. F.; KIRST, M. 2012. Accuracy of genomic selection methods in a standard dataset of loblolly pine (*Pinus taeda* L.). **Genetics**, 190:1503-1510.

RESENDE, M.D.V. 1991. Correções nas expressões do progresso genético com seleção em função da amostragem finita dentro de famílias e populações e implicações no melhoramento florestal. **Boletim de Pesquisa Florestal**, 22/23:61-77.

RESENDE, M.D.V. 1995. Delineamento de experimentos de seleção para a maximização da acurácia seletiva e progresso genético. **Revista Árvore**, 19:479-500.

RESENDE, M.D.V. 1997. Avanços da genética biométrica florestal. In: BANDEL, G.; VELLO, N.A.; MIRANDA FILHO, J.B. (Ed.). **Encontro sobre temas de genética e melhoramento: genética biométrica vegetal**. Anais, Esalq, Piracicaba. p.20-46.

RESENDE, M.D.V. 1999. **Predição de valores genéticos, componentes de variância, delineamentos de cruzamento e estrutura de populações no melhoramento florestal**. Universidade Federal do Paraná, Curitiba.

RESENDE, M.D.V. 2000a. **Análise estatística de modelos mistos via REML/BLUP no melhoramento de plantas perenes**. Embrapa Florestas, Colombo.

RESENDE, M.D.V. 2000b. **Inferência bayesiana e simulação estocástica (amostragem de Gibbs) na estimação de componentes de variância e valores genéticos em plantas perenes**. Embrapa Florestas, Colombo.

RESENDE, M.D.V. 2002. **Genética biométrica e estatística no melhoramento de plantas perenes**. Emprapa Informação Tecnológica, Brasília.

RESENDE, M.D.V. 2004. **Métodos Estatísticos Ótimos na Análise de Experimentos de Campo**. Embrapa Florestas, Colombo.

RESENDE, M.D.V. 2007a. **Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético**. Embrapa Florestas, Colombo. 561 p.

RESENDE, M.D.V. 2007b. Selegen-Reml/Blup: Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos. Embrapa Florestas, Colombo.

RESENDE, M.D.V. 2015. **Genética Quantitativa e de Populações**. Suprema, Visconde do Rio Branco.

RESENDE, M.D.V. 2016. Software Selegen-REML/BLUP: a useful tool for plant breeding. **Crop Breeding and Applied Biotechnology**, 16:330-339.

RESENDE, M.D.V., RESENDE JR., M.F.R., SANSALONI, C.; PETROLI, C.; MISSIAGGIA, A. A.; AGUIAR, A. M.; ABAD, J.I.M.; TAKAHASHI, E.; ROSADO, A.M.; FARIA, D.; PAPPAS, G.; KILIAN, A.; GRATTAPAGLIA, D. 2012b. Genomic Selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist**, 194:116-128.

RESENDE, M.D.V.; AZEVEDO, C.F.; SILVA, F.F.; NASCIMENTO, M.; GOIS, I.B.; ALVES, R.S. 2018. **Modelos Hierárquicos Generalizados Lineares Mistos (HGLMM), Máxima Verossimilhança Hierárquica (HIML) e HG-BLUP: Unificação das Três Classes de Inferência (Frequentista, Fisheriana e Bayesiana) na Análise Estatística em Biometria e Genética**. Suprema, Visconde do Rio Branco.

RESENDE, M.D.V.; BARBOSA, M.H.P. 2006. Selection via simulated BLUP based on family genotypic effects in sugarcane. **Pesquisa Agropecuária Brasileira**, 41:421-429.

RESENDE, M.D.V.; DUARTE, J.B. 2007. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical**, 37:182-194.

RESENDE, M.D.V.; DUDA, L.L.; GUIMARÃES, P.R.B.; FERNANDES, J.S.C. 2001c. Análise de modelos lineares mistos via Inferência Bayesiana. **Revista de Matemática e Estatística**, 21:41-70.

RESENDE, M.D.V.; FERNANDES, J.S.C. 1999a. Procedimento BLUP individual para delineamentos experimentais aplicados ao melhoramento florestal. **Revista de Matemática e Estatística**, 17:89-107.

RESENDE, M.D.V.; FERNANDES, J.S.C. 2000. Análises alternativas envolvendo o procedimento BLUP e o delineamento experimental de blocos incompletos ou látice. **Revista de Matemática e Estatística**, 18:103-124.

RESENDE, M.D.V.; FERNANDES, J.S.C.; SIMEÃO, R.M. 1999b. BLUP individual multivariado em presença de interação genótipos x ambientes para delineamentos experimentais repetidos em vários ambientes. **Revista de Matemática e Estatística**, 17:209-228.

RESENDE, M.D.V.; FURLANI-JUNIOR, E.; MORAES, M.L.T.; FAZUOLI, L.C. 2001a. Estimação de parâmetros genéticos e predição de valores genotípicos no melhoramento do cafeeiro pelo procedimento REML/BLUP. **Bragantia**, 60:185-193.

RESENDE, M.D.V.; HIGA, A.R. 1994. Maximização da eficiência da seleção em testes de progênies de *Eucalyptus* através da utilização de todos os efeitos do modelo matemático. **Pesquisa Florestal Brasileira**, 28/29:37-55.

RESENDE, M.D.V.; HIGA, A.R.; LAVORANTI, O.J. 1993. Predição de valores genéticos no melhoramento de *Eucalyptus* - melhor predição linear (BLP). In: **Congresso Florestal Brasileiro**, Anais, SBS, Curitiba. p. 144-147.

RESENDE, M.D.V.; OLIVEIRA, E.B.; HIGA, A.R. 1990. Utilização de índices de seleção no melhoramento do *Eucalyptus*. **Pesquisa Florestal Brasileira**, 21:1-13.

RESENDE, M.D.V.; PRATES, D.F.; JESUS, A.; YAMADA, C.K. 1996. Estimação de componentes de variância e predição de valores genéticos pelo método da máxima verossimilhança restrita (REML) e melhor predição linear não viciada (BLUP) em *Pinus*. **Pesquisa Florestal Brasileira**, 32/33:18-45.

RESENDE, M.D.V.; RAMALHO, M.A.P.; CARNEIRO, P.C.S.; CARNEIRO, J.E.S; BATISTA, L.G.; GOIS, I.B. 2016.Selection index with parents, populations, progenies and generations effects in autogamous plant breeding. **Crop Science**, 56:530-546.

RESENDE, M.D.V.; RAMALHO, M.A.P.; GUILHERME; S.; ABREU, A.F.B. 2015. Multigeneration index in the within progenies bulk method for breeding of self-pollinated plants. **Crop Science**, 55:1202-1211.

RESENDE, M.D.V.; REZENDE, G.D.S.P.; FERNANDES, J.S.C. 2001b. Regressão aleatória e funções de covariância na análise de medidas repetidas. **Revista de Matemática e Estatística**, 19:21-40.

RESENDE, M.D.V.; SILVA, F.F.; LOPES, P.S.; AZEVEDO, C.F. 2012a. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. Universidade Federal de Viçosa, Viçosa.

RESENDE, M.D.V.; SILVA, F.F; AZEVEDO, C.F. 2014. **Estatística Matemática, Biométrica e Computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência.** Suprema, Visconde do Rio Branco.

RESENDE, M.D.V.; STRINGER, J.K.; CULLIS, B.C; THOMPSON, R. 2005. Joint modelling of competition and spatial variability in forest field trials. **Brazilian Journal of Mathematics and Statistics,** 23:7-22.

RESENDE, M.D.V.; STURION, J. A. 2003. Análise estatística espacial de experimentos via modelos mistos individuais com erros modelados por processos ARIMA em duas dimensões. **Revista de Matemática e Estatística**, 21:7-33.

RESENDE, M.D.V.; STURION, J.A. 2001. **Analise genética de dados com dependência espacial e temporal no melhoramento de plantas perenes via modelos geoestatísticos e de series temporais empregando REML/BLUP ao nível individual**. Embrapa Florestas, Colombo.

RESENDE, M.D.V.; THOMPSON, R. 2003. **Multivariate spatial statistical analysis of multiple experiments and longitudinal data**. Colombo: Embrapa Florestas.

RESENDE, M.D.V.; THOMPSON, R. 2004. Factor analytic multiplicative mixed models in the analysis of multiple experiments. **Revista de Matemática e Estatística,** 22:1-22.

RESENDE, M.D.V; HIGA, A.R.; LAVORANTI, O.J. 1994. Regressão geno-fenotípica multivariada e maximização do progresso genético em programas de melhoramento de *Eucalyptus*. **Boletim de Pesquisa Florestal**, 28/29:57-71.

RESENDE, M.D.V; LOPES, P.S.; SILVA, R.L.; PIRES, I.E. 2008. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira,** 56:63-78.

RESENDE, M.D.V; THOMPSON, R.; WELHAM, S.J. 2006. Multivariate spatial statistical analysis of longitudinal data in perennial crops. **Revista de Matemática e Estatística**, 24:147-169.

ROBERTSON, A. 1955. Prediction equations in quantitative genetics. **Biometrics**,11:95-98.

RONNINGEN, K. 1971. Some properties of the selection index derived by "Henderson's mixed model method". **Z. Tierz Zuchtungsbiol**, 88:186.

SANTOS, V.S.; MARTINS FILHO, S.; RESENDE, M.D.V.; AZEVEDO, C.F.; LOPES, P.S.; GUIMARÃES, S.E.F.; SILVA, F.F. 2016. Genomic prediction for additive and dominance effects of censored traits in pigs. **Genetics and Molecular Research**, 15: gmr15048764.

SANTOS, V.S.; MARTINS FILHO, S.; RESENDE, M.D.V.; AZEVEDO, C.F.; LOPES, P.S.; GUIMARÃES, S.E.F.; GLÓRIA, L.S.; SILVA, F.F. 2015. Genomic selection for slaughter age in pigs using the Cox frailty model. **Genetics and Molecular Research**, 14:12616-12627.

SCHWARZ, G. 1978. Estimating the dimension of a model. **Annals of Statistics**, 6:461-464.

SEARLE, S.R. 1971. A biometrics invited paper. Topics in variance component estimation. **Biometrics**, 27:1-76.

SNEDECOR, G.W.; COCHRAN, W.G. 1967. **Statistical methods**. Iowa State University Press, Iowa.

SORENSEN, D.; GIANOLA, D. 2002. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. Springer,Verlag.

SORENSEN, D.; WAAGEPETERSEN R. 2003. Normal linear models with genetically structured residual variance heterogeneity: a case study. **Genetics Research**, 82:207-222.

STONE, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. **Journal of the Royal Statistical Society**, 39:44-47.

STRAM, D.O.; LEE, J.W. 1994. Variance components testing in longitudinal mixed effects setting. **Biometrics**, 50:1171-1177.

THOMPSON, R. 1973. The estimation of variance and covariance components when records are subject to culling. **Biometrics**, 29:527-550.

THOMPSON, R. 2019. Desert island papers - a life in variance parameter and quantitative genetic parameters estimation reviewed using sixteen papers. **Journal of Animal Breeding and Genetics**, 136:230-242.

VAN RADEN, P.M. 2008. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, 91: 4414-4423.

VENCOVSKY, R. 1987. Herança quantitativa. In: PATERNIANI, E.; VIEGAS, G.P. (Ed.). **Melhoramento e produção de milho**. Fundação Cargill, 137-214.

VERBYLA, A.P. 2019. A note on model selection using information criteria for general linear models estimated using REML. **Australian & New Zealand Journal of Statistics**, 61:39-50.

VIANA, A.P.; RESENDE, M.D.V. 2014. **Genética Quantitativa no Melhoramento de Fruteiras**. Interciência, Rio de Janeiro.

VIANA, J.M.S.; ALMEIDA, Í.F.; RESENDE, M.D.V.; FARIA, V.R.; SILVA, F.F. 2010.BLUP for genetic evaluation of plants in non-inbred families of annual crops. **Euphytica**, 174:31-39.

VIANA, J.M.S.; ALMEIDA, R.V.; FARIA, V.R.; RESENDE, M.D.V.; SILVA, F.F. 2011a. Genetic evaluation of inbred plants based on BLUP of breeding value and general combining ability. **Crop & Pasture Science**, 62:515-522.

VIANA, J.M.S.; FARIA, V.; SILVA, F.F.; RESENDE, M.D.V. 2011b. Best linear unbiased prediction and family selection in crop species. **Crop Science**, 51:2371-2381.

VISSCHER, P.M.; HILL, W.G.; WRAY, N.R. 2008. Heritability in the genomics era: concepts and misconceptions. **Nature Reviews Genetics**, 9:255-266.

VISSCHER, P.M.; MEDLAND, S.E.; FERREIRA, M.A.R.; MORLEY, K.I.; ZHU, G.; CORNES, B.K; MONTGOMERY G.W.; MARTIN, N.G. 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. **PLoS Genetics**, 2:e41.

VISSCHER, P.M.; YANG, J.; GODDARD, M.E. 2010. A commentary on 'Common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). **Twin Research and Human Genetics**, 13:517-524.

VOLPATO, L.; ALVES, R.S.; TEODORO, P.E.; RESENDE, M.D.V.; NASCIMENTO, M.; NASCIMENTO, A.C.C.; LUDKE, W.H.; SILVA, F.L.; BORÉM, A. 2019. Multi-trait multi-environment models in the genetic selection of segregating soybean progeny. **PLoS ONE**, 14:e0215315.

WEARDEN, S. 1959. The use of the power function to determine an adequate number of progeny per sire in a genetic experiment involving half-sibs. **Biometrics**, 15:417-423.

WEBER, W.E. 1982. Selection in segregating of autogamous species. I. Selection response for combined selection. **Euphytica**, 31:493-502.

WELLER, J.I. 2016. **Genomic Selection in Animals**. John Wiley & Sons, Hoboken.

WHITTAKER, J.C.; THOMPSON, R.; DENHAM, M.C. 2000. Marker assisted selection using ridge regression. **Genetical Research**, 75:249-252.

WITKOVSKÝ, V. 2012. Estimation, testing, and prediction regions of the fixed and random effects by solving the Henderson's mixed model equations. **Measurement Science Review**, 12:6.

WRIGHT, J.W. 1962. **Genetics of forest tree improvement**. FAO, Rome.