CV- α : designing validation sets to increase the precision and enable multiple comparison tests in genomic prediction

Rafael Massahiro Yassue¹; José Felipe Gonzaga Sabadin¹; Giovanni Galli¹; Filipe Couto Alves²; Roberto Fritsche-Neto^{1*}

¹Department of Genetics, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, São Paulo, Brazil.

²Departments of Epidemiology and Biostatistics, Statistics and Probability and Institute of Quantitative Health Science and Engineering, Michigan State University, East Lansing, USA

Abbreviations: CV, cross-validation; CV- α , cross-validation alpha-based design; GP, Genomic prediction; RRS, Repeated Random Subsampling; TGV, true genetic value.

Received _____.

*Corresponding author (e-mail): roberto.neto@usp.br.

Abstract Usually, the comparison among genomic prediction models is based on validation schemes as Repeated Random Subsampling (RRS) or K-fold cross-validation. Nevertheless, the design of training and validation sets has a high effect on the way and subjectiveness that we compare models. Those procedures cited above have an overlap across replicates that might cause an overestimated estimate and lack of residuals independence due to resampling issues and might cause less accurate results. Furthermore, posthoc tests, such as ANOVA, are not recommended due to assumption unfulfilled regarding residuals independence. Thus, we propose a new way to sample observations to build training and validation sets based on cross-validation alpha-based design (CV- α). The CV- α was meant to create several scenarios of validation (replicates x folds), regardless of the number of treatments. Using CV- α , the number of genotypes in the same fold across replicates was much lower than K-fold, indicating higher residual independence. Therefore, based on the CV- α results, as proof of concept, via ANOVA, we could compare the proposed methodology to RRS and K-fold, applying four genomic prediction models with a simulated and real dataset. Concerning the predictive ability and bias, all validation methods showed similar performance. However, regarding the mean squared error and coefficient of variation, the CV- α method presented the best performance under the evaluated scenarios. Moreover, as it has no additional cost nor complexity, it is more reliable and allows the use of non-subjective methods to compare models and factors. Therefore, CV- α can be considered a more precise validation methodology for model selection.

KEYWORDS: Repeated random subsampling; K-fold; Cross-validation, model selection

Introduction

Genomic prediction (GP) proposed by Meuwissen et al. (2001) evolved over the years, but it aims to estimate breeding values of unevaluated genotypes. Hence, it is an important tool for plant breeders to shorten the breeding cycle, increase selection accuracy, and assess genetic variation (Heff et al. 2010; Crossa et al. 2017). Usually, to evaluate the prediction accuracy of the genomic prediction models, the data is divided into training and validation sets. The first set is used to fit the genomic prediction model and estimate the marker effects, whereas the validation set is used to validate the effects estimated in the training set and estimate the accuracy of the predictions (Crossa et al. 2011).

In the genomic prediction context, several methods and parameters have been proposed for the comparison of prediction models (Blondel et al. 2015). Nevertheless, the predictive ability and the bias of the measures are two of the most commonly utilized to evaluate the superiority and goodness of models and scenarios. The former is estimated by Pearson's correlation between the predicted and true breeding values of individuals contained in the validation set. The latter is obtained by regressing the predicted breeding values over the true ones to obtain the regression coefficient, which indicates the shrinkage (compression) between both (Piepho et al. 2008; Luan et al. 2009).

Some studies have shown that the model accuracy is influenced by the training and validation set (Akdemir et al. 2015; Wu et al. 2015; Auinger et al. 2016), being the main schemes to design training and validation sets in GP studies are K-fold cross-validation (Burgueño et al. 2012; Crossa et al. 2014; Fè et al. 2016) and Repeated Random Subsampling (RRS), also called Monte Carlo CV (Würschum et al. 2014; Yu et al. 2016; Zhang et al. 2016). The first consists of splitting the data into k groups (folds) and fit a model using each fold as training and validation sets. In this sense, if $k = 5$, the model will be fitted five times. The second consists of randomly split the dataset into training and validation sets. Both schemes are generally repeated n times (see Arlot and Celisse, 2010).

The accuracy estimate obtained by K-fold might be affected by the number of folds, fold size, and the number of replicates (Wong 2015). Likewise, in cross-validation schemes, the RRS is influenced by the relation between training and validation sets and the number of replicates (Kohavi

1995). Furthermore, some factors may lead to biased estimates of predictive ability, such as overlapping between the training and validation set and different relatedness between individuals through sets (Runcie e Cheng 2019). The overlap between training and validation sets over replicates may cause biased results due to the predictions be correlated and non-independent residuals (Amer e Banos 2010). Therefore, neither validation schemes guarantee independence among replicates due to resampling issues. Thus, researchers cannot use standard and non-subjective methods to compare models and factors, such as ANOVA and other multiple comparison tests, due to assumptions unfulfilled regarding residuals independence.

It is import point out that as the number of treatments increases, it becomes a challenge to design orthogonal training and validation sets across the replicates without increase substantially the number of replicates. This problem is similar to experimental field designs involving a large number of treatments. However, the balanced incomplete blocks design seeks to maintain homogeneity among blocks and orthogonality across replicates (Yates, 1936). These schemes are widely used to evaluate the quality of models and their selection for field experiments. Moreover, an extension of cross-validation (CV) schemes applying balanced incomplete block design was first proposed by Shao (1993), considering that each fold is treated as "block" and each genotype as a "treatment." The orthogonal distribution of the treatments across the blocks within replicates in the balanced incomplete block designs will guarantee that every pair of treatments appears together according to some rules. Therefore, the CV schemes using the incomplete block design may increase the quality of estimates (Fuchs e Krautenbacher 2016), residuals independence, and may allow further multiple comparison analyses.

Based on described above, in this study, we propose a new method to design the training and validation sets for genomic prediction studies based on an alpha-lattice design scheme, called cross-validation alpha-based design (CV- α) and compare its performance to the methods commonly applied in GP studies for model selection. Also, based on the CV- α results, a case of study, via analysis of

84 variance (ANOVA), we could compare the proposed methodology to RRS and K-fold, applying four
85 genomic prediction models with a simulated and real dataset.

86 **Material and Methods**

87 In order to demonstrate the properties of the proposed cross-validation scheme, we aimed to
88 mimic a standard genomic prediction study, for instance, comparing kernels and statistical methods.
89 Thus, our aim is not comparing genomic matrices or Bayesian and frequentists approaches but simply
90 show that our cross-validation scheme allows multiple comparison tests. For that, we create a
91 simulated population (knowing the true parameters) and also used a well-known real dataset.

93 *Simulated dataset*

94 We simulated a population of maize single-crosses from inbred parents to perform genomic
95 prediction studies. For this, we used the *AlphaSimR* package (Gaynor 2019). A founder population of
96 1,000 individuals was simulated with ten chromosomes containing 30,000 segregating loci (SNPs).
97 The individuals were inbred and diploid. Forty-nine individuals were randomly sampled and crossed
98 to compose a partial diallel to obtain 906 hybrids. The phenotypic value (adjusted mean based on
99 heritability) was simulated by randomly sampling 500 QTN from the segregating loci with mean 100
100 and variance 50. The narrow and broad-sense heritabilities were set to be equal to 0.30 and 0.50,
101 respectively. Finally, to understand the effect of the validation methods in the predictive ability and
102 bias of the true genetic (TGV) and phenotypic value, we performed genomics prediction using both
103 metrics. We repeated the simulations 25 times and averaged the estimates above.

105 *An empirical case of study: USP maize dataset*

106 We used a dataset of 906 maize single-crosses from a full diallel among 49 tropical inbred
107 lines, according to Griffing's method 4 (Griffing 1956). The experiments were evaluated in two
108 locations, two years, and under two nitrogen levels. The genotypic information from the 49 tropical
109 inbred lines was obtained from Affymetrix® Axiom® Maize Genotyping Array, containing about
110 614,000 SNPs (Unterseer et al. 2014). For more details about the phenotypic and genotypic data, see
111 (Fristche-Neto et al. 2018).

The markers with a lower call rate (< 95%), heterozygous loci on at least one individual, and linkage disequilibrium (> 0.90) were removed. The missing markers were imputed using the *Beagle* 4.0 algorithm (Browning e Browning 2009) from the *synbreed* R package (Wimmer et al. 2012). Later, the genotype of each hybrid was built by combining the genotypes of its parental lines and hybrids with minor allele frequency (MAF < 0.05) were removed. After quality control, a total of 32,207 SNPs was available for further analysis.

To perform the genomic prediction studies, we evaluated the grain yield (GY, Mg ha⁻¹), corrected to 13% moisture, and stand across the eight environments. It was used to estimate the BLUP for hybrids following the model:

$$\mathbf{y} = \mathbf{S}\mathbf{l} + \mathbf{X}\mathbf{b} + \mathbf{W}\mathbf{c} + \mathbf{T}\mathbf{g} + \mathbf{U}\mathbf{i} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is the vector of the phenotypic value of hybrids; $\mathbf{l}$ is the vector of fixed effects of the environment (the combination of site x year x N level); $\mathbf{b}$ is the vector of fixed effects of blocks within an environment; $\mathbf{c}$ is the vector of fixed effects of checks; $\mathbf{g}$ is genotypic values, where $\mathbf{g} \sim N(0, \mathbf{I}\sigma_g^2)$; $\mathbf{i}$ is the interaction between environments and checks, where $\mathbf{i} \sim N(0, \mathbf{I}\sigma_{ge}^2)$; $\boldsymbol{\varepsilon}$ is the vector of random residuals from checks and hybrid by environments effects, where $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$. σ_ε^2 was jointly estimated based on e environments with r replicated checks in each site. $\mathbf{S}$, $\mathbf{X}$, $\mathbf{W}$, $\mathbf{T}$, and $\mathbf{U}$ are the incidence matrices for $\mathbf{l}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{g}$, and $\mathbf{i}$ (Fristche-Neto et al. 2018).

Genomic prediction

To perform the genomic prediction, we used the additive GBLUP model and the Reproducing Kernel Hilbert Spaces regression (RKHS). The following model equation is the general form of these two approaches:

$$\hat{\mathbf{g}} = \mathbf{1}\mu + \mathbf{Z}\mathbf{a} + \boldsymbol{\varepsilon}$$

where $\hat{\mathbf{g}}$ is the vector of BLUP; μ is the intercept; $\mathbf{a}$ is the vector of additive genetic effects with $\mathbf{a} \sim N(0, \mathbf{G}\sigma_a^2)$; and $\boldsymbol{\varepsilon}$ is the vector of random residuals with $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$. $\mathbf{1}$ is the incidence vector of μ , and $\mathbf{Z}$ is the incidence matrix for $\mathbf{a}$. $\mathbf{G}$ is the genomic relationship matrix ($\mathbf{G}_a$ – additive

genomic relationship matrix, and $\mathbf{K}$ – for Gaussian kernel), and $\mathbf{I}$ is the identity matrix. σ_a^2 is the additive genetic variance for $\mathbf{G}_a$ or genetic variance for $\mathbf{K}$, and σ_e^2 is the residual variance. The additive genomic relationship matrix ($\mathbf{G}_a$) was calculated as $\mathbf{G}_a = \mathbf{W}\mathbf{W}' / 2 \sum_{i=1}^n p_i(1 - p_i)$, where $\mathbf{W}$ is the centered matrix of SNPs, and p_i is the frequency of the allele p in locus i (VanRaden 2008). The Gaussian kernel ($\mathbf{K}$) was calculated as $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-hd_{ij}^2/q_{0.05})$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the marker vectors for the i^{th} and j^{th} individuals, respectively, and $q_{0.05}$ is the fifth percentile for the squared Euclidean distance $d_{ij}^2 = \sum_k (x_{ik} - x_{jk})^2$ (Pérez-Elizalde et al., 2015). The h value was considered equal to 1.

We used fitted two prediction models considering two statistical approaches (frequentist and Bayesian Genomic Best Linear Unbiased Predictor - GBLUP), resulting in four scenarios: 1) GBLUP with $\mathbf{G}_a$ kernel (GA_MM); 2) GBLUP with $\mathbf{K}$ kernel (GK_MM); 3) Bayesian GBLUP with $\mathbf{G}_a$ kernel (GA_Bayes), and 4) Bayesian GBLUP with $\mathbf{K}$ kernel (GK_Bayes).

The analyses were performed using *ASReml-R* (Butler et al. 2009), and *BGLR* (Pérez and de los Campos, 2014) packages for R. For Bayesian GBLUP models were performed using 10,000 iterations, 3,000 burn-in, and 5 thinning values. The convergence checks for Bayesian models are available in the Supplemental Figure S1 and S2.

Cross-validation alpha-based design

The cross-validation alpha-based design (CV- α) is an extension of the methodology presented by (Shao 1993) and consists of assigning treatments to folds in each replication by applying the alpha-lattice sorting premises. The CV- α was intended to create scenarios with two, three, or four replicates, regardless of the number of treatments. Each replicate is split into folds, and the number of folds will determine the percentage of training and validation sets. Each fold across replicates is based on the $\alpha(0,1)$ lattice design aiming to reduce the concurrences of any two treatments in the same fold (block) across the replicates (Patterson e Williams 1976).

However, the $\alpha(0,1)$ -lattice design assumptions involve the number of blocks (s) and block

size (k) (number of the folds and fold size, in our context) to determine the number of treatments (Patterson e Williams 1976). As the number of treatments is variable in a real scenario, we compute the nearest smallest number to attend the assumptions above, and the remaining treatments are randomly allocated into the folds. The alpha lattice design was created using the *agricolae* package (Mendiburu 2019), and the scripts are available at Github (<https://github.com/allogamous/CV-Alpha>).

In order to compare CV- α with the other two benchmarks schemes, we simulated two scenarios: 5-folds with four replicates and 10-folds with two replicates. First, we simulated a scenario with the number of treatments varying from 200 to 2,000 and computed the percentage of remaining treatments that were randomly assigned into folds for each scenario. After, we compared the same two benchmarks schemes according to the mean and standard deviation of the concurrence of any two genotypes, i.e., the number of folds containing both genotypes. The simulations were replicated ten times.

Model comparison

To evaluate the cross-validation alpha-based design (CV- α) performance, we compared it to benchmark validation schemes: repeated random subsampling (RRS) and K-fold using real and simulated datasets for genomic prediction. For RRS, we used 100 replicates, each with 80% of the data for the training set, and the remaining 20% of the data for the validation set, whereas for CV- α and K-fold were used five-folds and four replicates. The number of replicates or folds for each method considers the most common values for genome prediction studies using Bayesian and frequentist approaches (Zhao et al. 2013; Zhang et al. 2015, 2016; Yu et al. 2016).

From those, we obtained the predictive ability of each statistical model for the different CV methods. The predictive ability was estimated as Pearson's correlations between the predicted and observed phenotypes. For each CV method, we estimated the slope coefficient for the regression of the predicted values of the validation sets on its phenotypes. For this, the regression coefficient between predicted and genetic value was considered the prediction bias, measuring the degree of

190 inflation/deflation of prediction genomics. Nonbiased models are expected to have a regression
 191 coefficient equal to 1. For CV- α and K-fold, the level of averaging considered was at replicates.
 192 Although RRS and K-fold schemes do not have independence between replicates, ANOVA have been
 193 used to compare the predictive abilities from different models, even breaking the independence
 194 assumption. To verify how variance components of models are affected by these methods, we perform
 195 the ANOVA test considering the following model:

$$196 \quad \mathbf{l} = \mathbf{1}\mu + \mathbf{X}_1\mathbf{m} + \mathbf{X}_2\mathbf{n} + \mathbf{X}_3\mathbf{o} + \boldsymbol{\varepsilon}$$

197 where $\mathbf{l}$ is the vector of Pearson correlation transformed by Fisher z-transformation using the
 198 R package *DescTools* (Signorell et al., 2019); μ is the overall mean; $\mathbf{m}$ is the vector of statistical
 199 approach effect; $\mathbf{n}$ is the vector of relationship kernel; $\mathbf{o}$ is the vector of interaction between statistical
 200 approach and kernel; and $\boldsymbol{\varepsilon}$ is the vector of residuals. $\mathbf{X}_1$, $\mathbf{X}_2$ e $\mathbf{X}_3$ are incidence matrices for $\mathbf{m}$, $\mathbf{n}$, and
 201 $\mathbf{o}$, respectively. Quadratic components were estimated by the method of moments based on mean
 202 square expectation.

203 Results

204 *CV- α*

205 We performed several analyses to evaluate cross-validation alpha-based design (CV- α)
 206 performance. For this, we computed the number of treatments that were randomly assigned among
 207 folds and the concurrence between pairs of treatments in the same fold across replicates (Figure 1).
 208 The results reveal that the proportion of treatments randomly assigned among folds reduces as the
 209 number of treatments increases and tends to converge to 0.38% and 0.27% for five-folds with four
 210 replicates and ten folds with two replicates, respectively (Figure 1). Considering the concurrence
 211 between pairs of treatments in the same fold across replicates, the CV- α reveals lower mean and
 212 standard deviation in both evaluated scenarios when compared with the K-fold CV (Figure 2).

213

214 *Genomic prediction (simulated dataset)*

215 To understand the effects of validation schemes on genomic prediction, we simulated
 216 populations to obtain true genetic values (TGV) and phenotypic values. The validation methods did
 217 not significantly influence the average prediction ability of TGV and phenotypic values.
 218 Nevertheless, the RRS has several “extreme” values when compared to K-fold and CV- α . Besides,
 219 RRS showed a more substantial variation for bias, with several values overtaking 0.5 and 1.5 for
 220 phenotypic and TGV. (Figure 3).

221 For PA and bias, TGV, and phenotypic value, in terms of mean and standard deviation, the
 222 three validation methods do not differ among them (Table 1), except for phenotypic bias for RRS. On
 223 the other hand, when we considered mean squared error (MSE) and coefficient of variation (CV),
 224 CV- α showed the lowest CV for all scenarios evaluated, when compared with RRS and K-fold.

225

226

227 *Proof of concept*

228 For the maize dataset, PA and bias showed similar mean values for all validation methods. In
 229 terms of SD, K-fold, and CV- α presented similar performance and were lower than RRS (Table 2).

230 For mean squared error and coefficient of variation, CV- α presented lower values than K-fold and
231 RRS. The coefficient of variation for K-fold was 34.70% and 10% higher than CV- α for PA and bias,
232 respectively (Table 2).

233 We applied the CV- α to validate two statistical approaches (Bayesian and Mixed models) and
234 two types of kernels (Additive and Gaussian kernel) for genomic prediction models (Table 3). For
235 this, we applied a two-way ANOVA, and it was observed significative effects for types of the kernel
236 for predictive ability and bias. Gaussian kernel (**K**) presented higher PA (0.44) than **G_a** (0.42) and
237 lower bias (1.00 and 0.98, for **K** and **G_a**, respectively). For the type of two statistical approaches, the
238 Bayesian reveals a more biased estimation (0.98) when compared with GBLUP (1.01) (Table 3).

239 We can note that the proportion of phenotypic variance explained variation by each source of
240 variation vary across validation schemes (Figure 3). PA and bias had similar performance across CV
241 schemes for residual variance but vary for other variances. The RRS presented higher residual
242 variance and lower variances due to model effects. For the interaction, K-fold showed higher values
243 for PA. CV- α presented lower proportions of residual variances and higher variance due to the kernel
244 and statistical approaches effects.

245

246 Discussion

247 The main advantages of considering the α -design instead of the balanced incomplete block
248 design (BICV) are the flexibility regarding the number of treatments and folds (Singh e Bhatia 2017),
249 reduce the concurrence between pairs of treatments, increase the quality of estimates (Fuchs e
250 Krautenbacher 2016) and residuals independence, allowing further multiple comparison analyses.
251 The α -design is widely used in plant breeding experiments as well as its ANOVA (Alam et al. 2017;
252 Ta et al. 2018; Galic et al. 2019). Based on this, in the context of genomic prediction, the flexibility
253 of the CV- α is a good alternative to compare genomic selection models.

254 Our results reveal that CV- α reduces the concurrence between pairs of treatments (genotypes)
255 in the same fold across replicates and its standard deviation when compared with the K-fold scheme
256 (Figure 2). The concurrence of any two treatments causes dependence among folds, and comparative
257 tests become less precise. Thus, the CV- α designs fold and replicates with few or non-concurrence
258 across folds, generating a more independent and better scheme for composing training and validation
259 sets in a genomic prediction context.

260 Comparison between CV- α , K-fold, and RRS must be pondered since they have a different
261 level of averaging and different numbers of replicates compared with RRS, although CV- α and K-
262 fold are equivalent (Wong 2015). RRS showed a higher number of outliers, probability as results of
263 the different levels of average. However, it is an internal procedure for the method. The strategy to
264 divide folds and replicates according to the alpha-lattice design, as we suggest into CV- α , permits we
265 consider as replicate level mean, similar to replicate the effect in the alpha-lattice design.

266 Moreover, the RRS showed a large variation in the estimates for PA and, especially, for
267 prediction bias. We expected values for bias around 1.0. However, the RRS showed several values
268 overtaking 0.5 and 1.5, which shows a considerable inflation/deflation on the estimates. These results
269 indicate that RRS is a less accurate method, mainly when we use few replicates.

270 Estimates more accurate combined with few replicates to run a CV scheme is desirable,
271 especially when we consider a large number of genotypes, which is common in plant and animal

272 breeding. In these cases, to compute the inverse matrix, the genomic relationship matrix is a
273 challenge, and several studies have been aiming this (Misztal et al. 2014; Misztal 2016). Based on
274 this, CV- α is a good alternative to design CV schemes and has a more precise estimative in a case
275 where the number of replicates is a limitation.

276 The simulated and real datasets reveal that CV- α had a similar performance to K-fold and
277 RRS when compared in terms of mean and standard deviation for predictive ability and bias. On the
278 other hand, when we consider in terms of MSE and coefficient of variation, CV- α has better
279 performance due to higher independence across replicates.

280 Traditionally, in genomic prediction studies, model comparison and selection are based on
281 subjective methods such as mean and standard deviation without a comparative test. Some studies
282 also considered ANOVA and other statistical tests. Although due to assumption unfulfilled regarding
283 residuals independence and our results, this is not be recommended. CV- α reveals the lesser
284 occurrence of pairs of genotypes in the same fold across replicates, causing a more precise estimative.
285 The CV- α methodology consists of applying $\alpha(0,1)$ lattice design to design the folds across replicates,
286 and because of this, it allows post hoc test to model comparison.

287 The results above indicate that CV- α had a more precise estimative trough the reduction of
288 coefficient of variation, and the variance components were better discriminated across the factors in
289 the two-way ANOVA. It reveals how the impact of folds design across each replicate shift the
290 proportion of the total variation explained by each model factor reducing the residual variance.
291 Furthermore, the ANOVA test using RRS and K-fold to compare the performance of different models
292 can produce mistake conclusions, since the estimative of variance components load bias. Therefore,
293 CV- α allows determining how much variation each model factor has and compares different genomic
294 selection models based on the ANOVA test and posthoc test. Furthermore, the use of CV- α does not
295 imply any additional computer cost or complexity in the validation process of model selection.

296 As proof of concepts, we applied the proposed methodology to exemplify model selection.
297 For the simulated and maize dataset, both do not show considerable differences across approaches

298 (GBLUP and Bayesian) and kernel type (Additive genomic and Gaussian kernel) for predictive
 299 ability. Although, for the maize dataset, the use of **K** kernel showed higher predictive ability than **G_a**.
 300 This result is expected since the **K** kernel captures additive and non-additive effects (Heslot et al.
 301 2012). For bias, mixed models showed less biased results. Although, comparison among these models
 302 is not the focus of these studies since they have already been extensively studied (Chen et al. 2014;
 303 Gota e Gianola 2014; Cuevas et al. 2017).

304 In the context of genomic prediction studies, there are other ways to design training and
 305 validation sets. The CV- α may be expanded for these cases to better designing training and test sets
 306 across replicates and environments, such as CV1 and CV2 schemes (Burgueño et al. 2012) and other
 307 multi-environment and multi-trait studies. Also, the CV- α may be applied in any other cross-
 308 validation studies to select models and verify as the model factors behave according to the different
 309 sources of variation.

310 **Conclusion**

311 This study showed that the CV- α method is a good alternative to design cross-validations folds
 312 and replicates, mainly when researchers want to compare genomic prediction models, increasing
 313 precision in the model estimative, and to unravel the model factors impact in the total variation. Even
 314 though there were no differences in the mean and standard deviation for predictive ability and bias,
 315 our proposal was more accurate in terms of the mean squared error and coefficient of variation.
 316 Another advantage of CV- α is that it does not require any additional cost regarding computing
 317 demand or complexity. Furthermore, CV- α allows using the non-subjective methods to compare
 318 models and factors, through ANOVA and other multiple comparison tests, such as Tukey and Scott-
 319 Knott.

320 **Acknowledgment:** This study was financed in part by the Coordenação de Aperfeiçoamento de
321 Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Conselho Nacional de
322 Desenvolvimento Científico e Tecnológico (CNPq).
323 **Conflict of interest:** The authors declare no conflict of interest.

324 **References**

- 325 Akdemir D, Sanchez JI, Jannink JL (2015) Optimization of genomic selection training populations
326 with a genetic algorithm. *Genet Sel Evol* 47:1–10 . doi: 10.1186/s12711-015-0116-6
- 327 Alam MA, Seetharam K, Zaidi PH, et al (2017) Dissecting heat stress tolerance in tropical maize
328 (*Zea mays* L.). *F Crop Res* 204:110–119 . doi: 10.1016/j.fcr.2017.01.006
- 329 Amer PR, Banos G (2010) Implications of avoiding overlap between training and testing data sets
330 when evaluating genomic predictions of genetic merit. *J Dairy Sci* 93:3320–3330 . doi:
331 10.3168/jds.2009-2845
- 332 Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. *Stat Surv*
333 4:40–79 . doi: 10.1214/09-SS054
- 334 Auinger HJ, Schönleben M, Lehermeier C, et al (2016) Model training across multiple breeding
335 cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.). *Theor*
336 *Appl Genet* 129:2043–2053 . doi: 10.1007/s00122-016-2756-5
- 337 Blondel M, Onogi A, Iwata H, Ueda N (2015) A Ranking Approach to Genomic Selection. *PLoS*
338 *One* 10:e0128570 . doi: 10.1371/journal.pone.0128570
- 339 Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-
340 phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84:210–
341 223 . doi: 10.1016/j.ajhg.2009.01.005
- 342 Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values
343 when modeling genotype $\times$ environment interaction using pedigree and dense molecular
344 markers. *Crop Sci* 52:707–719 . doi: 10.2135/cropsci2011.06.0299
- 345 Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) ASReml User Guide Release 3.0. 398
- 346 Chen L, Li C, Sargolzaei M, Schenkel F (2014) Impact of genotype imputation on the performance
347 of GBLUP and Bayesian methods for genomic prediction. *PLoS One* 9:1–7 . doi:
348 10.1371/journal.pone.0101544
- 349 Crossa J, Pérez-Rodríguez P, Cuevas J, et al (2017) Genomic Selection in Plant Breeding: Methods,
350 Models, and Perspectives. *Trends Plant Sci* 22:961–975 . doi: 10.1016/j.tplants.2017.08.011
- 351 Crossa J, Pérez P, de los Campos G, et al (2011) Genomic selection and prediction in plant
352 breeding. *J Crop Improv* 25:239–261 . doi: 10.1080/15427528.2011.558767
- 353 Crossa J, Pérez P, Hickey J, et al (2014) Genomic prediction in CIMMYT maize and wheat
354 breeding programs. *Heredity (Edinb)* 112:48–60 . doi: 10.1038/hdy.2013.16
- 355 Cuevas J, Crossa J, Montesinos-López OA, et al (2017) Bayesian genomic prediction with genotype
356 $\times$ environment interaction kernel models. *G3 Genes, Genomes, Genet* 7:41–53 . doi:
357 10.1534/g3.116.035584
- 358 Fè D, Ashraf BH, Pedersen MG, et al (2016) Accuracy of genomic prediction in a commercial
359 perennial ryegrass breeding program. *Plant Genome* 9: . doi:
360 10.3835/plantgenome2015.11.0110

361 Fristche-Neto R, Akdemir D, Jannink JL (2018) Accuracy of genomic selection to predict maize
362 single-crosses obtained through different mating designs. *Theor Appl Genet* 131:1153–1162 .
363 doi: 10.1007/s00122-018-3068-8

364 Fuchs M, Krautenbacher N (2016) Minimization and estimation of the variance of prediction errors
365 for cross-validation designs. *J Stat Theory Pract* 10:420–443 . doi:
366 10.1080/15598608.2016.1158675

367 Galic V, Franic M, Jambrovic A, et al (2019) Genetic correlations between photosynthetic and yield
368 performance in maize are different under two heat scenarios during flowering. *Front Plant Sci*
369 10:1–11 . doi: 10.3389/fpls.2019.00566

370 Gaynor C (2019) AlphaSimR: Breeding Program Simulations

371 Gota M, Gianola D (2014) Kernel-based whole-genome prediction of complex traits: A review.
372 *Front Genet* 5:1–13 . doi: 10.3389/fgene.2014.00363

373 Griffing B (1956) Concept of general and specific combining ability in relation to diallel crossing
374 systems. *Aust J Biol Sci* 9:463–493

375 Heff EL, Lorenz AJ, Jannink J, Sorrells ME (2010) Plant Breeding with Genomic Selection : Gain
376 per Unit Time and Cost. *Crop Sci* 50:1681–1690 . doi: 10.2135/cropsci2009.11.0662

377 Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: A
378 comparison of models. *Crop Sci* 52:146–160 . doi: 10.2135/cropsci2011.06.0297

379 Kohavi R (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model
380 Selection. *Int Jt Conf Artif Intell*

381 Luan T, Woolliams JA, Lien S, et al (2009) The Accuracy of Genomic Selection in Norwegian Red
382 Cattle Assessed by Cross-Validation. *Genetics* 1126:1119–1126 . doi:
383 10.1534/genetics.109.107391

384 Mendiburu F (2019) agricolae: Statistical Procedures for Agricultural Research

385 Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-
386 Wide Dense Marker Maps

387 Misztal I (2016) Inexpensive computation of the inverse of the genomic relationship matrix in
388 populations with small effective population size. *Genetics* 202:401–409 . doi:
389 10.1534/genetics.115.182089

390 Misztal I, Legarra A, Aguilar I (2014) Using recursion to compute the inverse of the genomic
391 relationship matrix. *J Dairy Sci* 97:3943–3952 . doi: 10.3168/jds.2013-7752

392 Patterson HD, Williams ER (1976) A new class of resolvable incomplete block designs. *Biometrika*
393 63:83–92 . doi: 10.1093/biomet/63.1.83

394 Pérez P, de los Campos G (2014) Genome-Wide Regression and Prediction with the BGLR
395 Statistical Package. *Genetics* 2:483–495

396 Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant
397 breeding and variety testing. *Euphytica* 161:209–228 . doi: 10.1007/s10681-007-9449-8

398 Runcie D, Cheng H (2019) Pitfalls and Remedies for Cross Validation with Multi-trait Genomic
399 Prediction Methods. *G3 Genes, Genomes, Genet* g3.400598.2019 . doi:
400 10.1534/g3.119.400598

401 Shao J (1993) Linear model selection by cross-validation. *J Am Stat Assoc* 88:486–494 . doi:
402 10.1016/j.jspi.2003.10.004

403 Signorell A (2019) DescTools: Tools for Descriptive Statistics

404 Singh P, Bhatia D (2017) Incomplete block designs for plant breeding experiments. *Agric Res J*
405 54:607–611 . doi: 10.5958/2395-146x.2017.00119.3

406 Ta KN, Khong NG, Ha TL, et al (2018) A genome-wide association study using a Vietnamese
407 landrace panel of rice (*Oryza sativa*) reveals new QTLs controlling panicle morphological
408 traits. *BMC Plant Biol* 18:1–15 . doi: 10.1186/s12870-018-1504-1

409 Unterseer S, Bauer E, Haberer G, et al (2014) A powerful tool for genome analysis in maize:
410 Development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics*
411 15:1–15 . doi: 10.1186/1471-2164-15-823

412 VanRaden PM (2008) Efficient methods to compute genomic predictions. *J Dairy Sci* 91:4414–
413 4423 . doi: 10.3168/jds.2007-0980

414 Wimmer V, Albrecht T, Auinger HJ, Schön CC (2012) Synbreed: A framework for the analysis of
415 genomic prediction data using R. *Bioinformatics* 28:2086–2087 . doi:
416 10.1093/bioinformatics/bts335

417 Wong TT (2015) Performance evaluation of classification algorithms by k-fold and leave-one-out
418 cross validation. *Pattern Recognit* 48:2839–2846 . doi: 10.1016/j.patcog.2015.03.009

419 Wu X, Lund MS, Sun D, et al (2015) Impact of relationships between test and training animals and
420 among training animals on reliability of genomic prediction. *J Anim Breed Genet* 132:366–375
421 . doi: 10.1111/jbg.12165

422 Würschum T, Abel S, Zhao Y (2014) Potential of genomic selection in rapeseed (*Brassica napus*
423 L.) breeding. *Plant Breed* 133:45–51 . doi: 10.1111/pbr.12137

424 Yates F (1936) Incomplete Randomized Blocks. *Ann Eugen* 7:121–140 . doi: 10.1111/j.1469-
425 1809.1936.tb02134.x

426 Yu X, Li X, Guo T, et al (2016) Genomic prediction contributing to a promising global strategy to
427 turbocharge gene banks. *Nat Plants* 2: . doi: 10.1038/nplants.2016.150

428 Zhang X, Pérez-Rodríguez P, Semagn K, et al (2015) Genomic prediction in biparental tropical
429 maize populations in water-stressed and well-watered environments using low-density and
430 GBS SNPs. *Heredity (Edinb)* 114:291–299 . doi: 10.1038/hdy.2014.99

431 Zhang X, Sallam A, Gao L, et al (2016) Establishment and optimization of genomic selection to
432 accelerate the domestication and improvement of intermediate wheatgrass. *Plant Genome* 9:1–
433 18 . doi: 10.3835/plantgenome2015.07.0059

434 Zhao Y, Zeng J, Fernando R, Reif JC (2013) Genomic prediction of hybrid wheat performance.
435 *Crop Sci* 53:802–810 . doi: 10.2135/cropsci2012.08.0463

Figure 1. The proportion of treatments randomly distributed into folds to attend the alpha-design presupposition using CV- α with 5-folds with four replicates (a), and 10-folds with two replicates (b), based on simulated data.

Figure 2. Concurrence (number of times that a pair of treatments appear together in the same fold) mean and standard deviation between treatments pairs in the same fold across replicates using CV- α and K-fold with 5 and 10 folds with 4 and 2 replicates, respectively, based on simulated data.

Figure 3. Predictive ability (PA) and bias for TGV and phenotypic value for three validation schemes (CV- α , K-fold, and RRS) and four genomic prediction models (scenarios).

Figure 4. The proportion of total variance decomposed into effects of the kernel, statistical approach, the interaction between the kernel and statistical approach, and residual for bias and predictive ability (PA) applied in three cross-validation schemes (CV-α, K-fold, RRS).

455

456 **Table 1.** Averaged of 25 simulated datasets for mean, standard deviation (SD), mean squared error
457 (MSE), and coefficient of variation (CV) for predictive ability (PA) and bias for three CV schemes
458 (CV- α , K-fold, and RRS)

Scheme	Parameter	Mean	SD	MSE	CV (%)
CV- α	PA of phenotypic value	0.331	0.063	0.00017	3.78
	PA of TGV	0.748	0.053	0.00022	1.50
	Phenotypic Bias	1.024	0.064	0.00217	4.26
	TGV Bias	1.049	0.149	0.00040	1.68
K-Fold	PA of phenotypic value	0.331	0.062	0.00016	3.84
	PA of TGV	0.748	0.053	0.00023	1.52
	Phenotypic Bias	1.027	0.072	0.00251	4.36
	TGV Bias	1.050	0.151	0.00049	1.80
RRS	PA of phenotypic value	0.331	0.084	0.00414	19.41
	PA of TGV	0.748	0.062	0.00616	8.10
	Phenotypic Bias	1.024	0.274	0.07283	25.47
	TGV Bias	1.050	0.192	0.01366	10.26

459

460

461 **Table 2.** Summary of ANOVA, mean, standard deviation (SD), and coefficient of variation (CV) for
462 three validation schemes (CV- α , K-fold, and RRS) for predictive ability (PA) and bias

Model	CV- α				K-fold				RRS					
	Df	PA	Bias	Df	PA	Bias	Df	PA	Bias	Df	PA	Bias		
		MS			MS			MS						
St.Approaches	1	0.0002	0.0049	*	1	0.0002	0.0048	*	1	0.0035	0.1278	.		
Kernel	1	0.0016	**	0.0025	.	1	0.0007	0.0005	.	1	0.1292	**	0.4748	**
St.Approaches:Kernel	1	0.0001		0.0001		1	0.0005	0.0002		1	0.0002		0.0123	
Residuals	12	0.0001		0.0007		12	0.0002	0.0009		396	0.0046		0.0340	
CV (%)		2.16		2.70			2.91		2.97		14.61		18.38	
Mean		0.433		0.99			0.440		1.02		0.433		1.00	
SD		0.014		0.033			0.016		0.033		0.070		0.188	

463 **, *, ., ns: Significant at 1%, 5% , 10% and non-significant of error probability by F- test.

464 Statistical approaches (St. Approaches), Degrees of freedom (Df), Predictive ability (PA), Mean Squared (MS)

465

Table 3. Means, marginal means, and Tukey's test for the type of kernels and statistical approaches for predictive ability (PA) and bias

	PA			
	G_a	K	Marginal Means	
Bayesian	0.424	0.436	0.430	
Mixed models	0.426	0.445	0.436	
Marginal Means	0.425	b	0.441	a
	Bias			
Bayesian	0.963	0.992	0.977	B
Mixed models	1.002	1.023	1.012	A
Marginal Means	0.982	b*	1.008	a

*Means followed by the same lowercase letter in the row and uppercase letter in the column do not differ by the Tukey test at 5% and 10% * probability.