MGIDI: a novel multi-trait index for genotype selection in plant breeding

Tiago Olivoto[1,*] Maicon Nardino[2]

[1] Centro Universitário UNIDEAU, Rua Jacob Gremmelmaier, 215, 99900-000, Getúlio Vargas, RS, Brasil

[2] Department of Agronomy, Federal University of Viçosa, 36570-000, Viçosa, MG, Brazil

[*] Correspondence: Tiago Olivoto <tiagoolivoto@gmail.com>

Abstract

The efficiency in genotype selection in plant breeding programs could be greater if based on multiple traits, but identifying genotypes that combine high performance across many traits has been a challenger task for breeders. Here, we introduced the theoretical foundations, validated the potential, provided the needed tools for future implementations, and showed how breeders can use the novel multi-trait genotype-ideotype distance index (MGIDI) for genotype selection in plant breeding programs. We illustrated the application of the method using data from a preliminary wheat trial with 14 traits assessed in 44 genotypes. The MGIDI provided negative gains ($-2.29\% \leq$ gains $\leq -0.07\%$, a total of -5.15%) for all the four traits that wanted to decrease and positive gains ($0.46\% \leq$ gains $\leq 14.28\%$, a total of 49.64%) for nine of the 10 traits that wanted to increase. The MGIDI should be useful for breeders who aim at the selection of genotypes based on multiple traits because it provides a unique and easy-to-interpret selection process that incorporates the underlying correlation structure of the data and doesn't require any arbitrary weighting coefficient.The use of the MGIDI will save tremendous time and resources in plant breeding trials of any crop, guaranteeing long-term gains in grain yield without jeopardizing genetic gains in other important traits, thus helping breed of better varieties fast enough to satisfy the growing global population needs for food. A step-by-step guide and an R package containing all the needed functions are provided to facilitate the implementation of the MGIDI index in future studies.

Key-words: genotype selection multi-trait selection, metan, plant breeding, R software

1 Introduction

Widely established over the past 7,000 yrs., agriculture has dramatically transformed human evolution. Crop improvements have been helping us to satisfy the challenge of fuelling a global population that has been growing quickly and can achieve more than 10 billion by 2100 (Azosea et al., 2016). Plant breeders have developed an impressive range of techniques to obtain crops with increased yield and better quality (Hickey et al., 2019). In 58 years (1961-2018), for example, the world total production of cereals increased by 262% –ranging from 9.87×10^8 to 3.58×10^9 t– while the cultivated area increased by only 12.2% in the same period (see Supplementary Fig. S1 online). These unparalleled increases are explained by the increment of on average 51 kg ha^{-1} yr^{-1} in yielding (FAOSTAT, 2020). For wheat (*Triticum aestivum* L.), the world's top crop by area harvested, studies in Southern Brazil (Follmann et al., 2017) revealed genetic gains for grain yield between 1.82% (41.54 kg ha^{-1} yr^{-1}) and 4.38% (84.13 kg ha^{-1} yr^{-1}), reflecting the important efforts of wheat breeding programs carried out in such location.

Experienced breeders often keep in mind a set of plant traits that, if brought together in one new genotype would lead to high performance. Here, this target genotype is treated as *ideotype*. The concept of *ideotype* was introduced nearly a half-century ago by Donald (1968), which defined an *ideotype* as *a plant model that is expected to yield a greater quantity or quality of the grain, oil, protein, or other useful product when developed as a cultivar*. For wheat, for example, breeders often look for healthily- and high-yielding plants, that provide grains with high baking quality. The idea behind ideotype-design is increasing crop performance, of course, focusing on the selection of genotypes based on multiple traits simultaneously. The challenge that still remains is then how we can we breed better varieties fast enough to satisfy the growing global population needs for food.

Several linear selection indexes (Cerón-Rojas and Crossa, 2018) help breeders to select superior genotypes. A widely-used base linear phenotypic selection index (Bhering et al., 2012;

Bizari et al., 2017; Burdon and Li, 2019; Jahufer and Casler, 2015) is the Smith-Hazel (SH) index (Hazel, 1943; Smith, 1936). To compute the SH index, breeders use the phenotypic and genotypic variance-covariance matrices as well as a vector of economic weights to determine how a vector of index coefficients has to be chosen to maximize the correlation of unknown genetic values and phenotypic values. Because the SH index requires inverting a phenotypic covariance matrix (Smith, 1936), the presence of multicollinearity –which will certainly appear when several traits are assessed– can result in poorly conditioned matrices and biased index coefficients, thus, affecting the estimates of genetic gains.

Besides the multicollinearity issue, breeders often face difficult choices to express the economic value of traits, converting them into realistic economic weightings (Bizari et al., 2017). To circumvent this issue, some studies use multiple iterations with different sets of economic weights to identify the optimum set associated with index coefficients that identify genotypes with superior genetic worth to achieve the desired genetic gain (Jahufer and Casler, 2015). This approach, however, may limit the use of SH index by a coding layman, mainly because –as far as we know– this procedure (of multiple iterations) is not implemented in any free and widely-used software.

The selection in advanced stages of breeding programs is often made in multi-environment trial data (Dalló et al., 2019; Jarquin et al., 2020; Tiago Olivoto, Alessandro D. C. Lúcio, et al., 2019; Woyann et al., 2020). The differences among environments in the level of expression of genotypic differences can seriously erode expected economic gain in the SH index if it is not addressed by setting appropriate economic weights for the expressions of a trait in the different environments (Burdon and Li, 2019). Fortunately, recent advances such as the multi-trait stability index (Tiago Olivoto, Alessandro D. C Lúcio, et al., 2019) have been proven to be useful for selecting genotypes based on both mean performance and stability (Zuffo et al., 2020).

Even though the SH index is widely used as a multi-trait selection index in plant breeding

trials, there has been evidence (Bhering et al., 2012) that the adoption of this index is not interesting for plant breeding. Since the combination of multivariate techniques is efficient to account for the multicollinearity issue in multi-trait indexes (Rocha et al., 2018), there would seem to be value in an investigation to develop an index that covers the weaknesses of the SH index in which all the traits are selected favorably and with satisfactory gains for the application in breeding programs (Bermudez and Pinheiro, 2020).

In light of this, we propose a novel multi-trait index based on factor analysis and genotype-ideotype distance (MGIDI), mainly focused on the selection of superior genotypes in plant breeding programs where multiple traits have been assessed. To validate the potential of the MGIDI index, we used data on a preliminary trial stage with 44 wheat genotypes and 14 assessed traits. The application of the MGIDI index in future studies is facilitated by a step-by-step guide and an R package containing all needed functions.

2 Materials and Methods

2.1 Field experiments

The field experiment was conducted during the 2019 growing season in the experimental area of the Department of Agronomy of the Federal University of Viçosa (UFV), Viçosa, MG, Brazil. We used 35 homozygote tropical wheat lines developed by the UFV Wheat Breeding Program, in addition to nine commercial genotypes, namely BRS 264, BRS 394, BRS 254 (Embrapa), TBio Duque, TBio Aton, TBio Sintonia, TBio Ponteiro, TBio Sossego (Biotrigo Genetica) and CD 1303 (Coodetec).

The genotypes were arranged in a randomized complete block design with three replications. The experimental plot consisted of 5-five m long cultivar rows spaced at 0.20 m, with a final population density of 350 plants m^{-2}.

2.2 Crop management

Basic fertilization was carried out according to the interpretation of the chemical analysis of the soil aiming to meet the crop requirements. In the sowing furrow, 300 kg ha^{-1} of the formula 08-28-16 (NPK) was applied. Topdressing fertilization consisted of 90 kg ha^{-1} of nitrogen (N) were applied in two stages; 50% at the beginning of tillering and 50% at the pseudostem erection, which matches the decimal stages Z21 and Z30 of the Zadoks (Zadoks et al., 1974) scale, respectively.

The chemical control of weeds was performed approximately 20 days after sowing using 3 g ha^{-1} a.i. of metsulfurom methyl. The chemical control of aphids (*Metopolophium dirhodum* and *Sitobion avenae*) was performed using 75 g ha^{-1} a.i. of acetamiprid at the post-anthesis, Z68 stage of Zadoks (Zadoks et al., 1974) scale. No fungicide application was performed aiming at identifying healthier genotypes.

2.3 Data collection

Data on 14 agronomic traits were collected from the useful area (three central cropping rows) of each plot: days to flowering (FLO, days), considering the period between Z10 (first leaf through the coleoptile) and Z65 of Zadoks (Zadoks et al., 1974) scale, in which 50% of the plants in the plot were in anthesis; Plant height (PH, cm), measured from the collection to the top of the spike, excluding the awns. Spike height (SH, cm), measured from the base of the plant to the height of the ear; Flag leaf height (FLH, cm), measured from the base of the plant to the flag leaf; disease note (DIS), assigned according to the severity of diseases in leaves and spikes, ranging from 5 (healthy plants without diseases) to 1 (plants with high disease severity); grain yield (GY kg ha^{-1}), determined by manually harvesting the three central cropping rows, adjusting to 13% moisture and converting the grain weight to the hectare scale; hectoliter weight (HW, kg hL^{-1}), determined by weighing a known volume (250 mL) of a sample using a Dalle-Molle scale, and transforming the result into the standard

unit. In the post-harvest period, 10 plants were randomly harvested from the experimental useful plot for measuring the following spike-related traits: number of spikelets per spike (NSS), number of grains per spikelets (NGS), spike length (SL, cm), spike weight (SW, g), number of grains per spike (NGE, n), grain mass per spike (GMS, g) and harvest index of the spike (HIE, GMS/SW). The selection aimed at selecting genotypes with lower values of FLO, PH, EH, and FLH, and higher values of DIS, GY, HW, NSE, NGS, EL, EW, NGE, GME, and HIEP.

2.4 Statistical analysis

2.4.1 Variance component analysis

Each trait was analyzed using a mixed-effect model based on the Eq. (1).

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is an $n[= \sum_{j=1}^{r}(gr)] \times 1$ vector of response variable, i.e., the response of the ith genotype in the jth block ($i = 1, 2, \ldots, g$; $j = 1, 2, \ldots, r$; $\mathbf{y} = [y_{11}, y_{12}, \ldots, y_{gr}]'$); $\mathbf{b}$ is an $1 \times r$ vector of unknown and unobservable fixed effects of block $\mathbf{b} = [\gamma_1, \gamma_2, ..., \gamma_r]'$; $\mathbf{u}$ is an $m[= 1 \times g]$ vector of unknown and unobservable random effects of genotype $\mathbf{u} = [\alpha_1, \alpha_2, ..., \alpha_g]'$; $\mathbf{X}$ is an $n \times r$ design matrix of 0s and 1s relating $\mathbf{y}$ to $\mathbf{b}$; $\mathbf{Z}$ is an $n \times m$ design matrix of 0s and 1s relating $\mathbf{y}$ to $\mathbf{u}$; and $\mathbf{e}$ is an $n \times 1$ vector of random errors $\mathbf{e} = [y_{11}, y_{12}, \ldots, y_{gr}]'$. Random vectors $\mathbf{u}$ and $\mathbf{e}$ are assumed to be normaly and independently distributed with zero mean and variance-covariance matrices $\mathbf{G}$ and $\mathbf{R}$ respectively, such that

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right) \tag{2}$$

where $\mathbf{G} = \hat{\sigma}_{\alpha}^2 \mathbf{I}_g$ and $\mathbf{R} = \hat{\sigma}_{\epsilon}^2 \mathbf{I}_n$, being $\hat{\sigma}_{\alpha}^2$ and $\hat{\sigma}_{\epsilon}^2$ the variances for genotypes and random errors, respectively; $\mathbf{I_g}$, and $\mathbf{I_n}$ are the identity matrices of order g, and n, respectively.

Variance components (**G** and **R**) were estimated using REstricted Maximum likelihood (REML) using an Expectation-Maximization algorithm (Dempster et al., 1977).

The significance of the genotype effects was tested by a likelihood ratio (LR) test, which for the model, compares the $-2(Res)log$ likelihoods for two models, one with the random term (full model) and other without the random term (reduced model). The probability was then obtained by a two-tailed χ-square test with one degree of freedom. The variance components generated from the analysis were used to estimate the broad-sense heritability (h^2) on a genotype mean basis (equation 3).

$$h^2 = \frac{\hat{\sigma}^2_\alpha}{\hat{\sigma}^2_\alpha + \frac{\hat{\sigma}^2_\epsilon}{r}} \tag{3}$$

Mean cross-products were calculated and resolved to estimate the genetic variance-covariance matrix that was later used in the estimation of the Smith-Hazel index.

This mixed model analysis was used to generate Best Linear Unbiased Predictors (BLUP) for each genotype, which were then used to construct a two-way table (**X**) consisting of 44 rows (one per genotype) and 14 columns (one per trait). This table was later used to compute the multi-trait indexes.

The mixed model analysis was performed in the R 4.0.0 software using the functions `gamem()` and `get_model_data()` of the package `metan` (Olivoto and Lúcio, 2020). The complete code can be found as Supplementary item S1.3.

2.4.2 The multi-trait genotype-ideotype distance index (MGIDI)

- Rescaling the traits

The first step to compute the MGIDI index was to rescale the matrix **X** so that all the values have a 0-100 range (Tiago Olivoto, Alessandro D. C. Lúcio, et al., 2019). The rescaled value for the jth trait of the ith genotype ($r\mathbf{X}_{ij}$) was obtained as described in Eq. (4).

$$r\mathbf{X}_{ij} = \frac{\eta_{nj} - \varphi_{nj}}{\eta_{oj} - \varphi_{oj}} \times (\theta_{ij} - \eta_{oj}) + \eta_{nj} \tag{4}$$

Where η_{nj} and φ_{nj} are the new maximum and minimum values for the trait j after rescaling, respectively; η_{oj} and φ_{oj} are the original maximum and minimum values for the trait j, respectively, and θ_{ij} is the original value for the jth trait of the ith genotype.

The values for η_{nj} and φ_{nj} were chosen as follows. For the traits in which lower values are desired (FLO, PH, EH, and FLH), we used $\eta_{nj} = 0$ and $\varphi_{nj} = 100$. For the traits in which higher values are desired (DIS, GY, HW, NSE, NGS, EL, EW, NGE, GME, and HIEP), we used $\eta_{nj} = 100$ and $\varphi_{nj} = 0$. After the rescaling procedure, a two-way table of rescaled values ($r\mathbf{X}$) was obtained. Each column of $r\mathbf{X}$ has a 0-100 range that considers the desired sense of selection (increase or decrease) and maintains the correlation structure of the original set of variables. An example of this rescaling procedure made with the function `resca()` is given in Supplementary material S1.4.1.1.

- Factor analysis

The second step was to compute an exploratory factor analysis to group correlated traits into factors and then estimate the factorial scores for each genotype. The factor analysis was based on Eq. (5).

$$\mathbf{X} = \mu + \mathbf{Lf} + \varepsilon \tag{5}$$

where $\mathbf{X}$ is a $p \times 1$ vector of observations; μ is a $p \times 1$ vector of standardized means; $\mathbf{L}$ is a $p \times f$ matrix of factorial loadings; $\mathbf{f}$ is a $p \times 1$ vector of common factors; and ε is a $p \times 1$ vector of residuals, being p and f the number of traits and common factors retained, respectively. The eigenvalues and eigenvectors were obtained from the correlation matrix of the two-way table $r\mathbf{X}$. The initial loadings were obtained considering only factors with eigenvalues higher

than one. The *varimax* (Kaiser, 1958) rotation criteria was used for the analytic rotation and estimation of final loadings. The scores for genotypes were obtained according to Eq. (6).

$$\mathbf{F} = \mathbf{Z}(\mathbf{A^T R^{-1}})^{\mathbf{T}} \tag{6}$$

where $\mathbf{F}$ is a $g \times f$ matrix with the factorial scores; $\mathbf{Z}$ is a $g \times p$ matrix with the standardized means ($r\mathbf{X}$); $\mathbf{A}$ is a $p \times f$ matrix of canonical loadings, and $\mathbf{R}$ is a $p \times p$ correlation matrix between the traits. g, f, and p represents the number of genotypes, factors retained, and analyzed traits, respectively.

- Ideotype planning

The third step of the MGIDI computation was ideotype planning. By definition (equation 4), the ideotype has the highest rescaled value (100) for all analyzed traits. Thus, the ideotype was defined by a $1 \times p$ vector $\mathbf{I}$ such that $\mathbf{I} = [100, 100, ..., 100]$. The scores of the ideotype were also estimated according to Eq. (6).

The fourth and last step was the estimation of the multi-trait genotype-ideotype distance index (MGIDI) according to Eq. (7).

$$MGIDI_i = \left[\sum_{j=1}^{f} (\gamma_{ij} - \gamma_j)^2\right]^{0.5} \tag{7}$$

where $MGIDI_i$ is the multi-trait genotype-ideotype distance index for the ith genotype; γ_{ij} is the score of the ith genotype in the jth factor ($i = 1, 2, ..., g$; $j = 1, 2, ..., f$), being g and f the number of genotypes and factors, respectively; and γ_j is the jth score of the ideotype. The genotype with the lowest MGIDI is then closer to the ideotype and therefore presents desired values for all the analyzed traits.

The proportion of the MGIDI index of the ith genotype explained by the jth factor (ω_{ij}) was computed as described in the Eq. (8).

$$\omega_{ij} = \frac{\sqrt{D_{ij}^2}}{\sum_{j=1}^{f} \sqrt{D_{ij}^2}} \tag{8}$$

where D_{ij} is the distance between the ith genotype and the ideotype for the jth factor. For a given genotype, factors with low contributions suggests that such a genotype is close to the ideotype for the traits within that factor.

The selection gain in percentage, SG(%), was calculated for each trait considering a selection intensity of 15% based on the following equation.

$$SG(\%) = \frac{(\bar{X}_s - \bar{X}_o) \times h^2}{\bar{X}_o} \times 100 \tag{9}$$

where $\bar{X}_s$ is the mean of the selected genotypes, $\bar{X}_o$ is the mean of the original population and h^2 is the broad-sense heritability defined in Eq. (3).

2.4.3 Comparing MGIDI with other multi-trait indexes

The Smith-Hazel (SH) classic index (Hazel, 1943; Smith, 1936), and the recently proposed FAI-BLUP index (Rocha et al., 2018) were computed with the functions `Smith_Hazel()` and `fai_blup()`, respectively, of the package `metan` (Olivoto and Lúcio, 2020) and were used to validate the potential of the MGIDI index in terms of genetic gains. The selection gains for SH and FAI-BLUP indexes were computed according to Eq. (9) considering a selection intensity of 15%.

For the FAI-BLUP index, the desired ideotype (DI) was defined by a vector with `"min"` for the traits in which lower values are desired and `"max"` for the traits in which higher values are desired. Then the distances from each genotype according to ideotypes were estimated and converted into spatial probability as shown in Eq. (10).

$$P_{ij} = \frac{\frac{1}{d_{ij}}}{\sum\limits_{i=1;j=1}^{i=n;j=m} \frac{1}{d_{ij}}} \tag{10}$$

where P_{ij} is the probability of the ith genotype ($i = 1, 2, ..., n$) to be similar to the jth ideotype ($j = 1, 2, ..., m$); d_{ij} is the genotype-ideotype distance from the ith genotype to the jth ideotype based on standardized mean Euclidean distance.

The SH index was computed according to the Eq. (11)

$$\mathbf{b} = \mathbf{P^{-1}Gw} \tag{11}$$

where $\mathbf{P}$ and $\mathbf{G}$ are phenotypic and genetic variance-covariance matrices, respectively, and $\mathbf{b}$ and $\mathbf{w}$ are vectors of index coefficients and economic weightings, respectively. All traits were considered as having the same economic weight ($\mathbf{w} = 1$) with negative sign (-1) for traits in which lower values are desired and positive sign (1) for traits in which higher values are desired, thus, enabling to place all traits data in the desired direction for selection (Smiderle et al., 2019).

The genetic worth ($\mathbf{I}$) of the individual genotype based on the evaluated traits were then computed as described in Eq. (12).

$$\mathbf{I} = \mathbf{Xb} \tag{12}$$

where $\mathbf{X}$ is an $g \times p$ matrix of BLUPs for the g genotypes and p traits and $\mathbf{b}$ is a $1 \times p$ vector of index coefficient for the p traits.

Two scenarios were considered. In the first one (SH-1), the SH index was computed with all the 14 traits. In the second one (SH-2), the multicollinearity-generating traits were excluded and the index was computed with the remaining traits. The traits to be excluded were chosen with the function `non_collinear_vars()` of the package `metan` (Olivoto and

Lúcio, 2020) so that there are no VIF great than 10 (Olivoto et al., 2017).

To quantify the coincidence of genotype's selection (CI), we computed the coincidence index (Hamblin and Zimmermann, 1986) between each pair of indexes as follows.

$$CI = \frac{A - C}{M - C} \times 100 \tag{13}$$

where A is the number of selected genotypes common to different methods; M is the number of genotypes selected according to the selection intensity ($M = 7$); and C is the number of expected genotypes selected by chance ($M \times 0.15$). The codes to compute all the procedures of this section can be found as Supplementary code S1.5.3 online.

3 Results

3.1 Variance components

The likelihood ratio (LR) test indicated highly significant genotype effects ($p < 0.005$) for all analyzed traits (Table 1). Except for GY, NGS, NGSP, and SL, all the other traits had the genotypic variance ($\hat{\sigma}_g^2$) as the main component of the phenotypic variance ($\hat{\sigma}_p^2$). The broad-sense heritability on a genotype mean basis (h^2) ranged from 0.51 (NGSP) to 0.92 (FLO). High values of heritability ($h^2 > 0.8$) were observed for DIS, FLH, FLO, GMS, HW, SH, and SW, suggesting goods prospects of selection gains for these traits.

3.2 Loadings and factor delineation

Five principal components were retained, which explained 87.02% of the total variation among the traits (Table 2). After *varimax* rotation, the average communality (h) was 0.87 (HW $0.76 \leq h \leq 0.98$ SH), indicating that a high proportion of each variable's variance was explained by the factors.

Table 1: Deviance analysis, estimated variance components and genetic parameters for 14 agronomic traits evaluated in 44 wheat genotypes

Traits[a]	$\hat{\sigma}_g^2(\%)$	$\hat{\sigma}_r^2(\%)$	$\hat{\sigma}_p^2$	h^2	CV_g	CV_r	CV_g/CV_r	p-value
DIS	0.98(74.2)	0.34(25.8)	1.32	0.90	34.84	20.53	1.70	1.36e-18
FLH	16.65(61.3)	10.5(38.7)	27.15	0.83	6.75	5.36	1.26	6.20e-12
GMS	0.06(59)	0.04(41)	0.10	0.81	14.74	12.30	1.20	5.18e-11
GY	157688.65(46.1)	184685.87(53.9)	342374.52	0.72	9.07	9.81	0.92	5.71e-07
HIS	8.47(55.3)	6.86(44.7)	15.33	0.79	3.90	3.50	1.11	1.08e-09
HW	4.47(57.3)	3.33(42.7)	7.80	0.80	2.77	2.39	1.16	2.19e-10
NGS	11.05(38.8)	17.45(61.2)	28.50	0.65	8.21	10.32	0.80	2.89e-05
NGSP	0.03(26.1)	0.07(73.9)	0.10	0.51	6.10	10.26	0.59	4.83e-03
NSS	1.12(52.1)	1.03(47.9)	2.15	0.77	6.80	6.53	1.04	1.15e-08
PH	16.89(54.7)	13.98(45.3)	30.87	0.78	4.75	4.32	1.10	1.66e-09
SH	20.44(60.4)	13.41(39.6)	33.86	0.82	5.81	4.71	1.23	1.47e-11
SL	0.45(40.2)	0.67(59.9)	1.13	0.67	7.80	9.52	0.82	1.45e-05
SW	0.11(63)	0.06(37.1)	0.17	0.84	15.07	11.56	1.30	1.29e-12

[a] DIS, disease; FLH, flag leaf height; FLO, flowering; GMS, grain mass per spike; GY, grain yield; HIS, harvest index of the spike; HW, hectoliter weight; NGS, number of grains per spike; NGSP, number of grains per spiklet; NSS, number of spiklets per spike; PH, plant height; SH, spike height; SL, spike length; SW, spike weight.

The 14 traits were grouped into the five factors (FA) as follows: FA1 the spike-related traits (NSS, SL, SW, NGS, and GMS); FA2 the traits related to grain weight (HW and HIS); FA3 the plant-related traits (PH, SH, and FLH); FA4 the traits FLO, DIS, and NGSP; and in the FA5 the GY (Table 2).

Selected genotypes and coincidence index

The genotypes selected by the MGIDI index were G12, G37, G18, G8, G11, G9, and G32 (Fig. 1a). G15 was very close to the cut point (red line that indicates the number of genotypes selected according to the selection pressure), which suggests that this genotype can present interesting features. Thus, the researcher should deserve attention to investigate genotypes that are very close to the cutpoint. Of the seven genotypes selected, the MGIDI index shares only four with the FAI-BLUP index, one with the SH-1 index, and three with the SH-2 index. Only one genotype (G32) was common to all indexes (Table 3).

Table 2: Eigenvalues, explained variance, factorial loadings after varimax rotation, and communalities obtained in the factor analysis

Trait[a]	FA1	FA2	FA3	FA4	FA5	h[b]
NSS	**-0.876**[c]	-0.260	0.055	-0.198	0.068	0.881
SL	**-0.843**	-0.204	0.294	0.118	0.115	0.865
SW	**-0.709**	0.257	0.392	0.473	-0.028	0.947
NGS	**-0.716**	-0.096	0.091	0.539	-0.086	0.827
GMS	**-0.578**	0.426	0.357	0.542	-0.010	0.937
HW	0.079	**0.847**	-0.177	0.051	0.040	0.759
HIS	0.541	**0.653**	-0.156	0.170	0.015	0.772
PH	-0.075	-0.171	**0.966**	0.013	-0.008	0.968
SH	-0.209	-0.192	**0.947**	0.032	0.011	0.978
FLH	-0.265	0.066	**0.900**	0.243	0.030	0.945
FLO	-0.124	0.013	0.400	**0.724**	0.355	0.826
DIS	0.142	0.563	0.034	**-0.617**	-0.294	0.805
NGSP	0.033	0.143	-0.007	**0.867**	-0.167	0.800
GY	0.033	-0.003	-0.005	-0.006	**-0.933**	0.872
Eigenvalues[d]	5.570	2.410	1.820	1.380	1.010	
Variance	39.760	17.230	12.990	9.840	7.200	
Accumulated	39.760	56.990	69.980	79.820	87.020	

[a] See Table 1 for the full trait names.

[b] h, Communality

[c] Bold values indicate the variables grouped within each factor.

[d] See the entirely list of eigenvalues as Supplementary item S1.4.1 online.

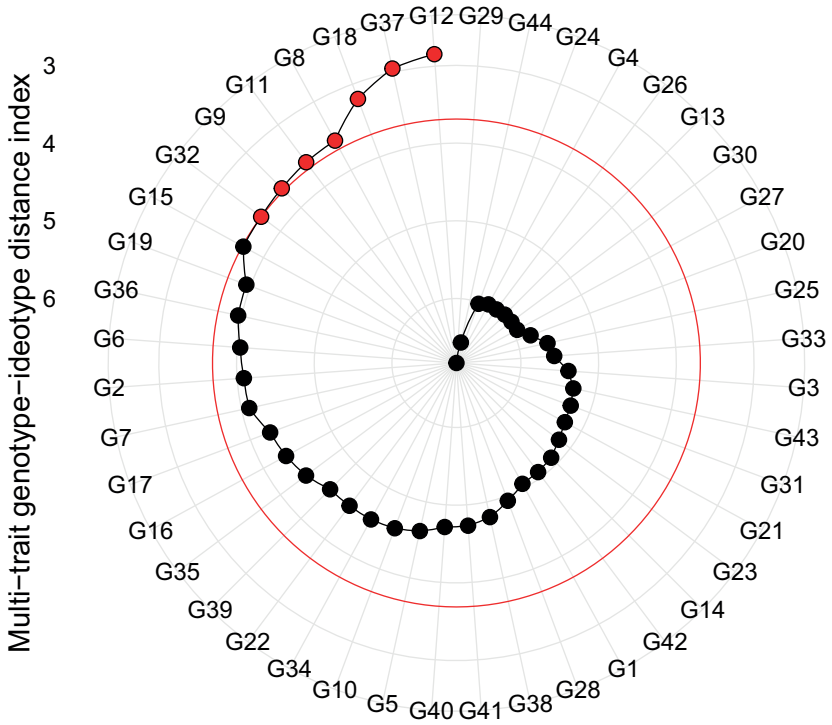

Figure 1: Genotype ranking in ascending order for the MGIDI index. The selected genotypes are shown in red color and the red circle represents the cutpoint according to the selection pressure. See the ranks for FAI-BLUP, SH-1, and SH-2 indexes as Supplementary Fig. S16.

Table 3: Coincidence index, number of common genotypes and shared genotypes for each pair of indexes evaluated

Index 1	Index 2	Coincidence	Common	Shared genotypes
MGIDI	FAI-BLUP	48.98	4	G12,G37,G18,G32
MGIDI	SH-1	-2.05	1	G32
MGIDI	SH-2	31.98	3	G18,G11,G32
FAI-BLUP	SH-1	14.96	2	G32,G7
FAI-BLUP	SH-2	48.97	4	G18,G32,G7,G2
SH-1	SH-2	31.97	3	G3,G7,G32

3.3 Predicted selection gains

Table 4 presents the comparison –via selection gains (SG)– between the studied indexes. The number of traits with desired gains was 13, 10, 5, and 9 for MGIDI, FAI-BLUP, SH-1, and SH-2, respectively. This suggests that MGIDI was the most efficient index to select genotypes with desired characteristics.

Table 4: Predicted genetic gains for the indexes MGIDI, FAI-BLUP, SH-1 and SH-2

Factor	Trait[a]	Goal	Genetic value	Genetic gain (%)			
				MGIDI	FAI-BLUP	SH-1	SH-2
FA1	NSS	Increase	15.54±0.14	2.22	-0.56	-4.05	-1.88
FA1	SL	Increase	8.63±0.08	3.86	-0.53	-4.71	-2.20
FA1	SW	Increase	2.17±0.04	14.28	9.92	-8.22	3.41
FA1	NGS	Increase	40.47±0.41	2.45	2.98	-3.16	1.00
FA1	GMS[b]	Increase	1.62±0.03	12.98	11.83	-6.22	4.18
FA2	HW	Increase	76.31±0.29	1.44	1.66	1.11	0.35
FA2	HIS	Increase	74.71±0.39	-0.55	2.13	1.71	1.08
FA3	PH	Decrease	86.46±0.55	-0.07	0.09	3.86	0.10
FA3	SH[b]	Decrease	77.83±0.62	-0.74	0.19	5.49	0.49
FA3	FLH	Decrease	60.42±0.56	-2.29	-1.37	4.89	-0.03
FA4	FLO	Decrease	60.45±0.52	-2.05	-2.45	4.40	1.11
FA4	DIS	Increase	2.84±0.14	8.54	4.50	20.66	11.23
FA4	NGSP	Increase	2.62±0.02	0.46	1.94	0.41	1.43
FA5	GY	Increase	4379.69±50.77	3.96	4.30	4.64	7.62
Total (positive)				49.64	38.17	2.17	26.22
Total (negative)				-5.15	-3.54	18.64	1.67

[a] See Table 1 for the full trait names.
[b] Traits removed from SH-2 due to multicollinearity problems.

The MGIDI index provided desired (positive) gains for all the spike-related traits within FA1 (Table 4). This was not observed for FAI-BLUP and SH-2 –which presented negative selection gains for NSS and SL–, and SH-1 –which presented negative selection gains for all traits in FA1.

The MGIDI index provided negative (-0.55%) gain for HIS, the only trait with undesired gains (Table 4). For WH, a positive gain was observed by the MGIDI index, however, with less magnitude compared to FAI-BLUP, SH-1, and SH-2 indexes (Table 4).

All the four traits in FA3 (plant-related traits) had desired (negative) gains using the MGIDI index. The FAI-BLUP, SH-1 and SH-2 indexes predicted positive gains for PH and SH. The gains for FLH considering the FAI-BLUP and SH-2 indexes were in the desired sense, but with lower magnitudes compared to the MGIDI index (Table 4).

The MGIDI and FAI-BLUP provided desired gains for all traits in FA4 (Table 4). SH-1 and SH-2 indexes provided higher gains for DIS compared to MGIDI and FAI-BLUP indexes, but the undesired (positive) gain for FLO indicates that the SH index was not efficient in selecting genotypes with desired gains for most of the traits.

The gains for GY –the single trait in FA5– was in desired sense (positive) for all the studied indexes. The magnitude of predicted gains followed the ascending order SH-2 > SH-1 > FAI-BLUP > MGIDI (Table 4). Even though MGIDI has provided lesser gain for GY, the total gains obtained with MGIDI presented the higher values, i.e., 49.62% for traits that wanted to increase and of -5.15% for traits that wanted to decrease. The gains for these same groups of traits for the FAI-BLUP index were of 38.16% and -3.55%, respectively. The SH index provided undesired gains when the multicollinearity was not considered (SH-1). In this case, the total gains were high and positive (18.64%) for traits that wanted to decrease and positive but low (2.17%) for traits that wanted to increase. After excluding the multicollienarity-generating traits (GMS and SH), the total gains were of 26.22% and 1.67% for traits that wanted to increase and decrease, respectively.

3.4 Contribution of factors to the MGIDI index

Presented in Fig. 2 is the proportion of each factor to the MGIDI index of the selected genotypes. This plot can be used to view the strengths and weaknesses of genotypes. For example, FA5 (GY) presented the smallest contribution for genotypes G32 and G11. This indicates that they were the most yielding genotypes among the selected ones (Fig. 3). On the other hand, FA5 was the most contributing factor for the MGIDI of G37, suggesting this genotype is poorly productive (Fig. 3). The genotype G3 was the most yielding genotype among the studied ones (Fig. 3), so, the smallest contribution of FA5 was for this genotype (Supplementary Fig. S3).

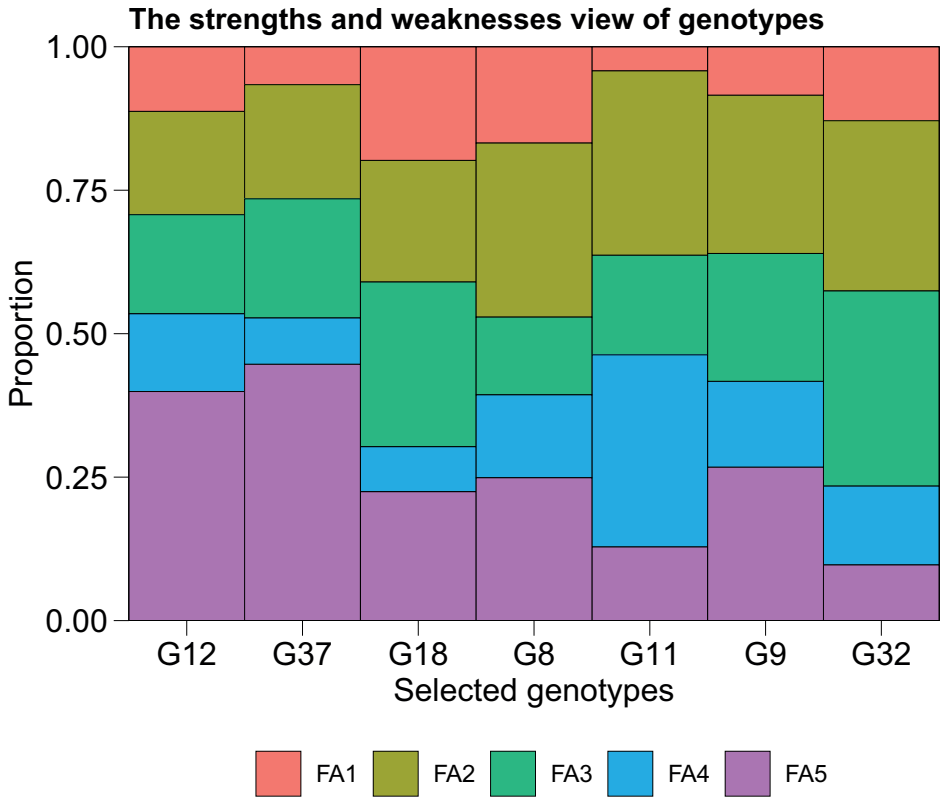

Figure 2: The strengths and weaknesses view of genotypes. The y-axis shows the proportion of each factor on the computed multi-trait genotype-ideotype distance (MGIDI) index of the selected genotypes. The smallest the proportion explained by a factor, the closer the traits within that factor are to the ideotype. The contribution of factors for all the studied genotypes can be found in Supplementary Fig. S3 online.

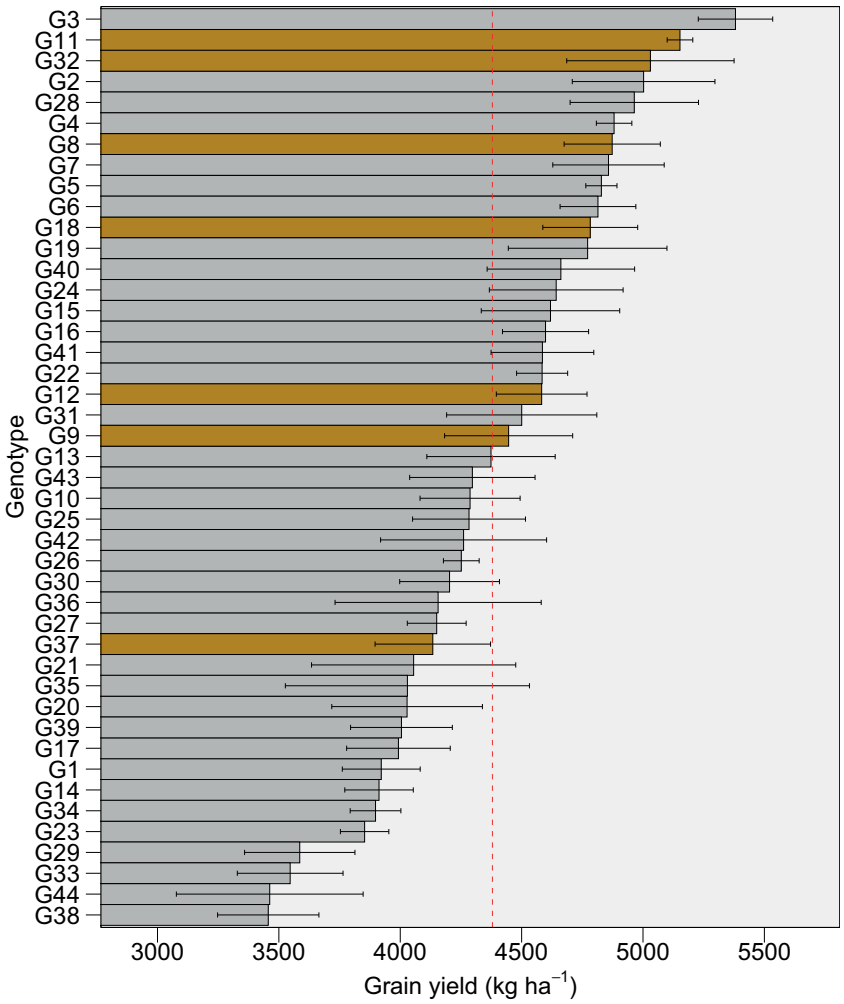

Figure 3: Grain yield for the 44 genotypes tested. The genotypes are sorted in increasing order of grain yield. Genotypes selected by the MGIDI index are shown in dark golden color. The dashed line shows the grand mean. Bars show the mean ±SE. N = 3.

The smallest contributions of FA1 (spike-related traits factor) was observed for G11 and G37. Since for all traits in FA1 higher values are desired, these genotypes should then have (simultaneously) high values for the traits within that factor. This is confirmed by looking at Supplementary Figs. S9 and S11-S14. The smallest contributions of FA2 for G12 and G37 (Fig. 2) suggests that these genotypes have high values of HW and HIS, compared to G11, for example, which has shown high contribution for FA2. The smallest contribution of FA3 (plant-related traits) for G8 (Fig. 2) indicates this genotype has a low stature (Supplementary

Figs. S4-S6). Since G8 also performed well regarding GY (Fig. 3), this genotype stands out as a potential genitor to obtain segregant populations with few stature and high grain yield. Finally, the smallest contributions of FA4 were observed for G18 and G37, indicating these genotypes have a shorter vegetative period (Supplementary Fig. S3) and above-average values for DIS (Supplementary Fig. S7) and NGSP (Supplementary Fig. S10).

4 Discussion

4.1 The theoretical basis of the MGIDI index

The MGIDI index uses the same rescaling procedure as the MTSI index (Tiago Olivoto, Alessandro D. C Lúcio, et al., 2019), which is used for genotype selection based on mean performance and stability. This rescaling procedure puts all traits in a 0-100 range, which facilitates the definition of the ideotype. In this case, an ideotype would have a value of 100 for all traits. This is only possible because the rescaling procedure considers the desired direction of selection (increase or decrease the trait values). In future studies, it will be up to the breeders to define the values of η_{nj} and φ_{nj} in Eq. (4) to rescale the traits.

In a multi-trait framework, the hypothesis is that the traits may be related in some way due to an underlying correlation structure that is unknown beforehand. In the MGIDI index, the factor analysis was used to account for this correlation structure, like in the FAI-BLUP index (Rocha et al., 2018). This was only possible because the rescaling procedure kept the original correlation structure of the data, allowing to plan an ideotype easily. Thus, it was possible to determine in how many latent variables the original set of traits could be reduced and the relationships among traits within latent variables. In our case, the 14 original traits were reduced into five final latent variables (factors) that explained ~87% of the original data. This dimensional reduction makes it simpler the interpretation and decision making by breeders. Because factor analysis produces orthogonal axes among final factors, it was possible to obtain genotype's scores free from multicollinearity. Finally, the distance from

a given genotype to the ideotype could be computed using the Euclidean distance. So, in contrast to conclusions of recent studies (Bermudez and Pinheiro, 2020), we have shown here that factor analysis is an efficient technique to establish an index in which most of the traits are selected favorably, showing satisfactory gains for the application in breeding programs.

The multicollinearity is a systemic issue in multivariate analyses (Graham, 2003; Olivoto et al., 2017), being a limiting issue in the SH index. Since the SH index requires inverting a phenotypic covariance matrix among traits (equation 11), the presence of highly correlated traits can result in either biased index coefficients –since $\mathbf{P}$ is not optimally conditioned– or in an infinite number of solutions if $\mathbf{P}$ is not positive definite. Even though non-collinear traits can be selected easily using the function `non_collinear_vars()` of the R package `metan` (Olivoto and Lúcio, 2020) –which would facilitate the implementation of the SH index in future studies– we strongly suggest the use of the MGIDI or even FAI-BLUP indexes to account for the multicollinearity in the data.

4.2 The MGIDI in practice

The MGIDI index was found to have many practical applications since it allows a unique and easy-to-interpret selection process. Besides dealing with collinear traits, the MGIDI index doesn't require the use of economic weights such as in the SH index, in which one can predict genetic and economic gains for various possible combinations of genetic parameters and assumed economic weights (Bizari et al., 2017; Burdon and Li, 2019). In our study, the MGIDI index was found to outperform the FAI-BLUP and SH indexes since it provided more balanced and in desired sense gains. Thus, MGIDI can help breeders to guarantee long-term gains in grain yield, without jeopardizing genetic gains in other important traits in wheat.

Functions to compute the MGIDI, FAI-BLUP, and SH indexes in R software have been elaborated and implemented in the R package `metan` (Olivoto and Lúcio, 2020). These functions will make it easier for the implementation of the MGIDI index in future studies.

The breeder uses a model fitted with the functions `gafem()` (fixed model) or `gamem()` (mixed model) as the input data in the function `mgidi()`, plans the ideotype by declaring which traits are desired to increase or decrease and the function will take care of the details. Since both FAI-BLUP and SH indexes are also implemented in `metan` (Olivoto and Lúcio, 2020) package, it will be easy –and we strongly suggest– the comparison of these indexes in terms of predicted genetic gains in future studies.

The proportion of MGIDI index explained by each factor (Fig. 2) is an important tool that can be used to identify the strengths and weaknesses of genotypes. It is possible to identify in the selected genotypes –or even in the non-selected ones (Supplementary Fig. S2)– traits or groups of traits that need to be improved. For example, the low contribution of FA5 in G11 (Fig. 2) indicates that this genotype is highly productive (Fig. 3) but was the selected genotype with lesser hectoliter weight (Supplementary Fig. S8) and lesser harvest index of the spike (Supplementary Fig. S15), which can be inferred due to the high contribution of FA2 (Fig. 2). G11 was also the healthier (Supplementary Fig. S7) and longer vegetative period (Supplementary Fig. S3) selected genotype, which is explained by the high contribution of FA4 for this genotype (Fig. 2). We may use then these contributions to chose possible genitors for future crossings. In our example, G8, G11, and G18 could be included in a crossing block aiming at obtaining plants with high production and healthy (G11), with a shorter vegetative period (G18), and with a lesser stature (G8). Compared to existing methods, this approach is graphical, objective, effective, and straightforward to identify the strengths and weaknesses of genotypes in plant breedings.

4.3 A step-by-step guide for future studies

Our supplementary portion (items 1.3 and 1.4) provides the codes used in this article that can be easily adapted in future studies. There have been two main ways to compute the MGIDI index. By using a two-way table of genotype-vs-trait mean or by using the raw data. In the last case, both fixed- and mixed-effect models can be considered.

To compute the MGIDI index based on a two-way table, breeders must provide a table with genotypes' labels as row names and variables in the columns (See an example in Supplementary material S1.5.1). In this case, since there is no information regarding the heritability of the traits, only the selection differential is computed.

When the raw data is available the MGIDI index can be computed from previously fitted models. The function `gafem()` is used to performs a genotype analysis using fixed-effect models. The function `gamem()` fits a mixed-effect model to the data –which was we have done in this article–. In this case, the genotype is assumed to have a random effect. An object of class gafem or gamem can be then used as input data in the `mgidi()` function to compute the MGIDI index (See the supplementary items S1.4 and S1.5.2). When a fitted model is used as input data, the heritability is extracted internally, allowing the computation of selection gains automatically.

Another useful function of the `metan` (Olivoto and Lúcio, 2020) package that can facilitate the implementation of the multi-trait indexes in future studies is `coincidence_index()` (Supplementary material S1.5.3). This function computes the coincidence index (Hamblin and Zimmermann, 1986) for each combination of indexes, thus, allowing us to compare the selected genotypes across different indexes (in our case MGIDI, FAI-BLUP, and SH) quickly.

5 Conclusion

In this study, we demonstrated how breeders can use the novel multi-trait genotype-ideotype distance index (MGIDI) for genotype selection in plant breeding trials. The potential of the MGIDI index was proved using real data coming from a preliminary wheat trial with 14 traits assessed in 44 genotypes. The MGIDI provided negative gains for all the four traits that wanted to decrease and positive gains for nine of the 10 traits that wanted to increase. In terms of total desired gains, the MGIDI index was found to outperform two widely used multi-trait indexes, namely, the Smith-Hazel and FAI-BLUP indexes. The MGIDI index

provides a unique selection process that is easy to interpret and considers the underlying correlation structure among the traits. Thus, MGIDI should be useful for breeders who aim at the selection of genotypes based on multiple traits, saving a lot of time and resources in plant breeding trials of any crop, thus helping breed of better varieties fast enough to satisfy the growing global population needs for food.

6 Data accessibility Statement

Additional supporting information may be found online in the Supporting Information section. The data set used to illustrate the method is publicly available at https://data.mendeley.com/datasets/p724jxkzmd/2 under a Creative Commons Attribution 4.0 International license.

7 Acknowledgements

We thank the graduate post-graduate students of the Federal University of Viçosa, who helped in the data collection and typing. We also thank the National Council for Scientific and Technological Development (CNPq) and Coordination for the Improvement of Higher Education Personnel (CAPES) for the scholarships to the students. The authors have no conflicts of interest to declare.

8 References

Azosea, J.J., Sevčíkováa, H., Rafterya, A.E., 2016. Probabilistic population projections with migration uncertainty. Proceedings of the National Academy of Sciences of the United States of America 113, 6460–6465. https://doi.org/10.1073/pnas.1606119113

Bermudez, F., Pinheiro, J.B., 2020. Selection to high productivity and stink bugs resistance by multivariate data analyses in soybean. Bragantia 79, 250–259. https://doi.org/10.1590/1678-4499.20190380

Bhering, L.L., Laviola, B.G., Salgado, C.C., Sanchez, C.F.B., Rosado, T.B., Alves, A.A., 2012. Genetic gains in physic nut using selection indexes. Pesquisa Agropecuaria Brasileira 47, 402–408. https://doi.org/10.1590/S0100-204X2012000300012

Bizari, E.H., Val, B.H.P., Pereira, E. de M., Mauro, A.O.D., Unêda-Trevisoli, S.H., 2017. Selection indices for agronomic traits in segregating populations of soybean. Revista Ciencia Agronomica 48, 110–117. https://doi.org/10.5935/1806-6690.20170012

Burdon, R.D., Li, Y., 2019. Genotype-environment interaction involving site differences in expression of genetic variation along with genotypic rank changes: simulations of economic significance. Tree Genetics & Genomes 15, 1–10. https://doi.org/10.1007/s11295-018-1308-3

Cerón-Rojas, J.J., Crossa, J., 2018. Linear selection indices in modern plant breeding. Springer International Publishing. https://doi.org/10.1007/978-3-319-91223-3

Dalló, S.C., Zdziarski, A.D., Woyann, L.G., Milioli, A.S., Zanella, R., Conte, J., Benin, G., 2019. Across year and year-by-year GGE biplot analysis to evaluate soybean performance and stability in multi-environment trials. Euphytica 215, 1–12. https://doi.org/10.1007/s10681-019-2438-x

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of The Royal Statistical Society, Series B 39, 1–38.

Donald, C.M., 1968. The breeding of crop ideotypes. Euphytica 17, 385–403. https://doi.org/10.1007/BF00056241

FAOSTAT, 2020. Food and Agriculture Organization of the United Nations.

Follmann, D.N., Cargnelutti Filho, A., Lúcio, A.D., Souza, V.Q. de, Caraffa, M., Wartha, C.A., 2017. Genetic progress in homogeneous regions of wheat cultivation in Rio Grande do Sul State, Brazil. Genetics and Molecular Research 16, gmr16019517. https://doi.org/10.4238/gmr16019517

Graham, M.H., 2003. Confronting Multicollinearity in Ecological Multiple Regression. Ecology 84, 2809–2815. https://doi.org/10.1890/02-3114

Hamblin, J., Zimmermann, M.J. de O., 1986. Breeding Common Bean for Yield in Mixtures, in: Plant Breeding Reviews. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 245–272. https://doi.org/10.1002/9781118061015.ch8

Hazel, L.N., 1943. The Genetic Basis for Constructing Selection Indexes. Genetics 28, 476–90.

Hickey, L.T., N. Hafeez, A., Robinson, H., Jackson, S.A., Leal-Bertioli, S.C.M., Tester, M., Gao, C., Godwin, I.D., Hayes, B.J., Wulff, B.B.H., 2019. Breeding crops to feed 10 billion. Nature Biotechnology 37, 744–754. https://doi.org/10.1038/s41587-019-0152-9

Jahufer, M.Z.Z., Casler, M.D., 2015. Application of the Smith-Hazel Selection Index for Improving Biomass Yield and Quality of Switchgrass. Crop Science 55, 1212–1222. https://doi.org/10.2135/cropsci2014.08.0575

Jarquin, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J.W.R., Covarrubias-Pazaran, G., Burgueño, J., Pacheco, A., Grondona, M., Wimmer, V., Prasanna, B.M., 2020. Genomic Prediction Enhanced Sparse Testing for Multi-environment Trials. G3: Genes, Genomes, Genetics g3.401349.2020. https://doi.org/10.1534/g3.120.401349

Kaiser, H.F., 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23, 187–200. https://doi.org/10.1007/BF02289233

Olivoto, T., Lúcio, A.D., 2020. metan: an R package for multi-environment trial analysis. Methods in Ecology and Evolution 11, 783–789. https://doi.org/10.1111/2041-210X.13384

Olivoto, T., Lúcio, A.D.C., Silva, J.A.G., Marchioro, V.S., Souza, V.Q., Jost, E., 2019. Mean Performance and Stability in Multi-Environment Trials I: Combining Features of AMMI and BLUP Techniques. Agronomy Journal 111, 2949–2960.

https://doi.org/10.2134/agronj2019.03.0220

Olivoto, T., Lúcio, A.D.C., Silva, J.A.G., Sari, B.G., Diel, M.I., 2019. Mean Performance and Stability in Multi-Environment Trials II: Selection Based on Multiple Traits. Agronomy Journal 111, 2961–2969. https://doi.org/10.2134/agronj2019.03.0221

Olivoto, T., Souza, V.Q., Nardino, M., Carvalho, I.R., Ferrari, M., Pelegrin, A.J., Szareski, V.J., Schmidt, D., 2017. Multicollinearity in path analysis: a simple method to reduce its effects. Agronomy Journal 109, 131–142. https://doi.org/10.2134/agronj2016.04.0196

Rocha, J.R. do A.S. de C., Machado, J.C., Carneiro, P.C.S., 2018. Multitrait index based on factor analysis and ideotype-design: proposal and application on elephant grass breeding for bioenergy. GCB Bioenergy 10, 52–60. https://doi.org/10.1111/gcbb.12443

Smiderle, É., Furtini, I., Botelho, F.B., Resende, M.P., Botelho, R.T., Colombari Filho, J.M., Castro, A.P. de, Utumi, M., 2019. Index selection for multiple traits in upland rice progenies. Revista de Ciências Agrárias 42, 1–10. https://doi.org/10.19084/RCA18059

Smith, H.F., 1936. A discriminant function for plant selection. Annals of Eugenics 7, 240–250. https://doi.org/10.1111/j.1469-1809.1936.tb02143.x

Woyann, L.G., Meira, D., Matei, G., Zdziarski, A.D., Dallacorte, L.V., Madella, L.A., Benin, G., 2020. Selection indexes based on linear-bilinear models applied to soybean breeding. Agronomy Journal 112, 175–182. https://doi.org/10.1002/agj2.20044

Zadoks, J.C., Chang, T.T., Konzak, C.F., 1974. A decimal code for the growth stages of cereals. Weed Research 14, 415–421. https://doi.org/10.1111/j.1365-3180.1974.tb01084.x

Zuffo, A.M., Steiner, F., Aguilera, J.G., Teodoro, P.E., Teodoro, L.P.R., Busch, A., 2020. Multi-trait stability index: A tool for simultaneous selection of soya bean

genotypes in drought and saline stress. Journal of Agronomy and Crop Science 00, 1–8. https://doi.org/10.1111/jac.12409