Multitrait index based on factor analysis and ideotype-design: proposal and application on elephant grass breeding for bioenergy

JOÃO ROMERO DO AMARAL SANTOS DE CARVALHO ROCHA, JUAREZ CAMPOLINA MACHADO and PEDRO CRESCÊNCIO SOUZA CARNEIRO

1Universidade Federal de Viçosa, Campus Universitário, Viçosa 36570-900, Brazil, 2Embrapa Gado de Leite, Rua Eugênio do Nascimento 610, Juiz de Fora 36038-330, Brazil

Abstract

This study proposes a new multitrait index based on factor analysis and ideotype-design (FAI-BLUP index), and validates its potential on the selection of elephant grass genotypes for energy cogeneration. Factor analysis was carried out, and afterwards, factorial scores of each ideotype were designed according to the desirable and undesirable factors, and the spatial probability was estimated based on genotype-ideotype distance, enabling genotype ranking. In order to quantify the potential of the FAI-BLUP index, genetic gains were predicted and compared with the Smith-Hazel classical index. The FAI-BLUP index allows ranking the genotypes based on multitrait, free from multicollinearity, and it does not require assigning weights, as in the case of the Smith-Hazel classical index and its derived indices. Furthermore, the genetic correlation - positive or negative - within each factor was taken into account, preserving their traits relationship, and giving biological meaning to the ideotypes. The FAI-BLUP index indicated the 15 elephant grass with the highest performance for conversion to bioenergy via combustion, and predicted balanced and desirable genetic gains for all traits. In addition, the FAI-BLUP index predicted gains of approximately 62% of direct selection, simultaneously for all traits that are desired to be increased, and approximately 33% for traits that are desired to be decreased. The genotypes selected by the FAI-BLUP index have potential to improve all traits simultaneously, while the Smith-Hazel classical index predicted gains of 66% for traits that are desired to be increased, and −32% for traits that are desired to be decreased, and it does not have potential to improve all traits simultaneously. The FAI-BLUP index provides an undoubtable selection process and can be used in any breeding programme aiming at selection based on multitrait.

Keywords: biomass, BLUP, Cenchrus purpureus, energy co-generation, multicollinearity, Napier grass, REML, selection index, SEM

Introduction

Elephant grass [Pennisetum purpureum Schum.; synonym Cenchrus purpureus (Schumach.) Morrone] may be used as forage or as bioenergy crop (Ra et al., 2012; Fontoura et al., 2015; Rocha et al., 2017b). Besides the morpho-agronomic traits, elephant grass breeders have focused on high quality nutrition attributes for animal feed, emphasizing high nitrogen concentration, low fibre levels and high digestibility (Rengsirikul et al., 2013). In turn, for bioenergy purposes, the selection of quality attributes occurs in the opposite direction, that is low nitrogen concentration and high fibre levels.

Experienced plant breeders keep in mind a plant ideotype that leads them to select high performance plants. In this context, the idea behind the ideotype is that it will provide to the breeders an ultimate target for selection, and thereby exchange the stepwise trial-and-error method, consequently increasing plant performance (Van Oijen & Höglin, 2016).

Selecting high performance genotypes for multitrait simultaneously can be a hard task. The first index for simultaneous selection was proposed by Smith (1936) for plant breeding, and by Hazel (1943), for animal breeding. This index is based on the selection of unknown genetic values. Thus, the use of phenotypic values and genetic covariances is necessary to determine how a vector of weights has to be chosen in order to maximize the correlation of unknown
genetic values and phenotypic values (Hazel et al., 1994).

Currently, the Smith-Hazel classical index has been successfully used. However, one of the difficulties of applying this index is the lack of a procedure to weight the traits of economic importance (Cerón-Rojas et al., 2006; Stephens et al., 2012). Some parameters have been assigned as relative economic weights, such as coefficient of genetic variation (Bhering et al., 2012) and heritability; in other cases, they can be randomly attributed (Zhang et al., 2009; Stephens et al., 2012).

In the selection process, plant breeders usually handle multitrait (Santchurn et al., 2012, 2014). However, multicollinearity problems will certainly appear and be another obstacle faced by the Smith-Hazel classical index. According to Prunier et al. (2015), multicollinearity between traits is a systemic issue in multivariate analyses, and is likely to cause serious difficulties to the proper interpretation of the results, with the risk of erroneous conclusions, misdirected research and inefficient conservation measures. Furthermore, the Smith-Hazel classical index does not take advantage of the genetic correlations between traits.

Several methods are designed to incorporate collinear variables, such as: factor analysis, principal component analysis, principal component regression, partial least squares, etc. Factor analysis may produce uncorrelated or orthogonal axes between final factors scores, and therefore they are free from multicollinearity (Dormann et al., 2013). This method concentrates in the first few new latent variables, and usually the less important latent variables are discarded, leading to dimensional reduction (Dormann et al., 2013), which consequently simplifies the analysis.

The theoretical foundations of structural equation modelling (SEM) arise by joining the traditional technique of factor analysis (Exploratory Factor Analysis) with the ideotype-design (Confirmatory Factor Analysis). SEM will allow the use of the correlations (covariance) between the traits. In this way, the influence that a trait exerts under another will be computed, getting information regarding their magnitude and sense.

In light of the aforementioned, this study proposes a new multitrait index based on factor analysis and genotype-ideotype distance (FAI-BLUP index), and validates the potential of FAI-BLUP index on the selection of elephant grass genotypes for conversion to bioenergy via combustion.

Materials and methods

Method description

Exploratory factor analysis. The principal components analysis was used to extract factorial loads of genetic correlation matrix, obtained by genetic values. The varimax criterion was used (Kaiser, 1958) for the analytic rotation and for the calculation of the factor scores of the weighted least squares method (Bartlett, 1938) was used.

Ideotypes-design (step by step). The number of ideotypes was defined based on the combination of desirable and undesirable factors for the objective of the selection. The number of ideotypes is given by the following algorithm:

\[ NI = 2^n, \]

in which:

\[ NI = \text{number of ideotypes}; \]

\[ n = \text{Number of factors}. \]

The number of factors (n) used to design the ideotypes should be equal to the number of eigenvalues (variance of principal components), higher or equal to one (Kaiser, 1958). It also indicates the number of coordinates that should be calculated (i.e. n factorial scores) for each ideotype. Thus, each ideotype and their descriptions based on the combination of desirable and undesirable factors are described as follows: If, for example, three eigenvalues are higher than one (n = 3), eight ideotypes (NI = 2^3 = 8) will be designed with three coordinates each one.

The factor score is a linear combination of standard genetic values (BLUP means) weighted by the canonical loadings obtained by the factor analysis. Therefore, a desirable factor should have the desirable genetic values (of the data set) for all traits under selection. The desirable genetic values may be the maximum, minimum, mean or a specific genetic value. An undesirable factor should have undesirable genetic values (maximum, minimum, mean or a specific genetic value) for all traits under selection. Thus, each ideotype and its coordinates (factorial scores) can be designed.

Multitrait index based on factor analysis and genotype-ideotype distance (FAI-BLUP index). After ideotypes were determined, the distances from each genotype according to ideotypes (genotype-ideotype distance) were estimated and converted into spatial probability, enabling the genotype ranking. The following algorithm was used:

\[ P_q = \frac{1}{\sum_{i=1}^{m} \frac{d_i}{d_q}}, \]

in which: \( P_q \) = Probability of the \( i^\text{th} \) genotype (\( i = 1, 2, \ldots, n \)) to be similar to the \( j^\text{th} \) ideotype (\( j = 1, 2, \ldots, m \)); \( d_q \) = Genotype-ideotype distance from the \( i^\text{th} \) genotype to the \( j^\text{th} \) ideotype – based on standardized mean Euclidean distance.

The theoretical basis of FAI-BLUP index. In addition to dealing very well with multicollinearity problems and the lack of weight assignment (in both cases due to factorial analysis), the FAI-BLUP index takes into account the correlation structure obtained from the data and directs to select genotypes closer to the hypothesized by the breeder (ideotype). The jointing of the exploratory factor analysis (EFA, i.e. sample by the data) and the ideotype-design (resembles confirmatory factor analysis – CFA) fit the theoretical basis of the structural equation modelling (SEM).
In EFA, the researcher has a large set of traits and hypothesizes that the observed traits may be linked together by virtue of the traits correlations (unknown underlying structure); the aim of an EFA is to uncover this structure (Ullman, 2006) and lead to dimensional reduction (Dormann et al., 2013).

In a CFA the researcher has a strong idea about the number of factors, the relations among the factors, and the relationship between the factors and measured traits; hence, the factor extraction and rotation are not part of CFA (Ullman, 2006). The goal of the CFA is to hypothesize a factor structured a priori and verified empirically (or by tests), rather than derived from the data (Ullman, 2006; Lei & Wu, 2007). However, in order to use this theoretical foundation in the breeding programmes selection process, the ideotype-design must be used, as it considers all the desirable relationships between the traits and the values for the traits that wish to achieve.

One of the main advantages of SEM is that it can be used to study the relationships among latent constructs (factors) that are indicated by multiple measures. It is also applicable to both experimental and non-experimental data, as well as cross-sectional and longitudinal data (Ullman, 2006; Lei & Wu, 2007). Structural equation modelling have been applied on multitrait genetic breeding by Gianola & Sorensen, 2004; Valente et al., 2010; Rosa et al., 2011 and Valente et al., 2013.

**Comparison of the FAI-BLUP index with the Smith-Hazel classical index.** The Smith-Hazel classical index was used to validate the potential of FAI-BLUP index by comparing both indices. The Smith-Hazel index aims at determining how a vector of weights has to be chosen to maximize the correlation of indices. The Smith-Hazel index considers all the desirable relationships between the traits and the values for the traits that wish to achieve.

One of the main advantages of SEM is that it can be used to study the relationships among latent constructs (factors) that are indicated by multiple measures. It is also applicable to both experimental and non-experimental data, as well as cross-sectional and longitudinal data (Ullman, 2006; Lei & Wu, 2007). Structural equation modelling have been applied on multitrait genetic breeding by Gianola & Sorensen, 2004; Valente et al., 2010; Rosa et al., 2011 and Valente et al., 2013.

**Method application**

**Experimental information.** The experiment was carried out at the experimental field of Embrapa Dairy Cattle Research Center, located in the municipality of Coronel Pacheco, MG, Brazil (21°33'18"S, 43°15'51"W, at 417 m asl), in a red-yellow latosol soil with the following chemical properties: pH (H2O) = 5.4; H+Al = 2.31 cmol dm⁻³; P (Mehlich 1) = 1.1 mg dm⁻³; K = 23 mg dm⁻³; and the following exchangeable cations (cmol dm⁻³): Al³⁺ = 0.2; Ca²⁺ = 1.4; Mg²⁺ = 0.7. The planting was carried out in December, 2010, in 0.20 m deep furrows, and 80 Kg ha⁻¹ P₂O₅ fertilizer was applied at planting. After the establishment stage, 30 days after planting, elephant grass was cut to 0.30 m stubble height (uniformity harvest). The first of the three growth periods started at this time. Maintenance fertilization was carried out with 300 Kg ha⁻¹ of the N-P₂O₅-K₂O formulation (20 : 05 : 20 blended granular fertilizer), after the uniformity harvest, and after the first evaluation cutting. Fertilization was carried out according to the soil analysis.

Three evaluation cuttings were carried out for this study. Aiming at using them as bioenergetic feedstock, 1st (28th September 2012) and 2nd (04th June 2013) cuttings were harvested at 250 days, and the 3rd (15th April 2014), at 315 days after the previous harvest.

**Genetic material and experimental design.** One hundred genotype of the Active Elephantgrass Germplasm Bank (BAGCE) were evaluated. Plots (1.5 m × 4 m) consisted of a single 4 m row. Rows were planted side by side, spaced 1.5 m apart. Plots were allocated in a simplex lattice design, with two replications.

**Evaluated Traits.** The following traits were measured: mean height (m) – obtained from the arithmetic mean of the height of three random plants samples, in each plot, measured from the ground level to the curve of the last completely expanded leaf; phenotypic vigor (1 to 5) – obtained using a grading scale, which ranged from 1 to 5 (5 = high vigor; 1 = low vigor); stalk diameter (mm) – obtained from the arithmetic mean of five plants in the useful plot, randomly taken, measured at 10 cm from the ground level with a digital caliper rule; total green biomass (Mg ha⁻¹) – obtained from a cut at 7.5 cm stubble height in a 3 m section from the middle of rows, using a gasoline powered trimmer, and then harvested by hand. The 3 m section was immediately weighed in the field to provide estimates of total green biomass. Total dry biomass (Mg ha⁻¹) was quantified by multiplying the total green biomass by the dry matter concentration (%).

Before cutting the experimental plots, random samples of three complete plants from each plot were collected. Then, these samples were preprocessed in a stationary forage cutter and dried in a forced air circulation oven at 56 °C for 72 h. After drying, samples were ground (1 mm) in a Wiley type grinder and sent to the biomass analysis laboratory for the chemical analysis described below:

- Acid detergent fibre (g Kg⁻¹), neutral detergent fibre (g Kg⁻¹), cellulose (g Kg⁻¹), lignin (g Kg⁻¹) and hemicellulose (g Kg⁻¹) – determined following the methodology proposed by Goering & Van Soest (1967); Cellulose/lignin ratio (−) – given by the cellulose/lignin ratio. In vitro digestibility of the dry biomass (g Kg⁻¹) – determined following the methodology used by Tilley & Terry (1963); Nitrogen (g Kg⁻¹) – stated following the methodology proposed by the Association of Official

© 2017 The Authors. GCB Bioenergy Published by John Wiley & Sons Ltd, 10, 52–60
Analytical Chemical (AOAC, 1975); Ash (g Kg⁻¹) – given according to the methodology proposed by Silva & Queiroz (2002); Calorific value (MJ kg⁻¹) – determined using an IKA C-5000 calorimeter; Dry matter concentration (g Kg⁻¹) – obtained by the sampling of three complete plants from each plot, which were dried in a kiln after weighing (fresh weight) until weight stabilization. Samples were weighed (dry weight) again, and then the dry matter concentration was determined by the ratio between dry weight and fresh weight. This trait was used as a common denominator for the estimation of cellulose, lignin, hemicellulose, in vitro digestibility of the dry biomass, nitrogen, ash and calorific value.

Statistical analyses

The mixed model methodology was adopted for statistical analyses via REML/BLUP (restricted residual maximum likelihood/best linear unbiased prediction), according to Patterson & Thompson (1971) and Henderson (1975).

The statistical model was denoted by:

\[ y = X m + Z g + W b + T i + Q p + e, \]

where; \( y \) = data vector; \( m \) = vector of the effects of the measurement-replication combination (assumed as fixed) added to the overall mean; \( g \) = vector of genetic effects (assumed as random); \( b \) = vector of block effects (assumed as random); \( i \) = vector of the genotype x measurements effects; \( p \) = vector of the permanent environment (random); \( e \) = vector of residue (random); \( X, Z, W, T \) and \( Q \) represent the incidence matrices for these effects.

For the random effects of the model, the significance for the likelihood ratio test was tested using the chi-square statistic with one degree of freedom. Genetic values (BLUP means) were predicted for each one of the 100 genotype based on the 16 traits evaluated.

Software

The software Selegen-REML/BLUP (Resende, 2007) was used for the deviance analysis, prediction of genetic values, genetic variance-covariance, genetic correlation and coefficient of genetic variation. The Smith-Hazel classical index was used on the software GENES (Cruz, 2013). The R software (R Development Core Team 2015) was used for the principal component analysis, factor analysis and genotype-ideotype distances (spatial probability). The FAI-BLUP-index routine analysis (Routine S1) applied to the R software is available in supporting information.

Results and Discussion

Genetic variability

Significant genotypes effect (\( P < 0.05 \)) was detected by the joint deviance analysis in the three cuts evaluated for 15 traits, with the exception of hemicellulose. As significant genetic variability among genotypes is essential for genetic progress, hemicellulose content will not be used in subsequent analyses.

Exploratory factor analysis

Table 1 shows the eigenvalues and cumulative frequency for the 15 principal components obtained by the genetic correlation matrix. Only the three-first principal components had eigenvalues higher than one, and thus, according to the Kaiser’s criterion (Kaiser, 1958), data may be condensed (dimensional reduction) in three factors. The cumulative frequency for the first three principal components, or communality mean (common variance) was higher than 78%, indicating that it is sufficient to represent 78% of all the variability (Table 1).

After varimax rotation (Table 2), high genetic correlation for the first factor was observed among the traits cellulose, lignin, cellulose/lignin ratio, acid detergent fibre, neutral detergent fibre, in vitro digestibility of the dry biomass, and nitrogen level, and this factor was named qualitmetric factor. For the second factor, high genetic correlation was observed among mean height, phenotypic vigor, total green biomass and total dry biomass, and it was named by phytometric factor. The third factor was named by energymetric factor, and stalk diameter, dry matter, ash and calorific value are strongly correlated. Genetic correlations among traits within a factor can be given in the same and/or opposite direction.

Idotype-design

To design the ideotype 1 (specific ideotype for direct combustion – Table 3), the maximum standard genetic values were used for the traits cellulose and lignin content, dry matter, calorific value, stalk diameter, mean height, phenotypic vigor, total green biomass, total dry biomass, and nitrogen level.

Table 1 Eigenvalue estimates by principal components analysis and the variance proportion explained by them.

| Eigenvalue | Eigenvalue (%) | Cumulative variance (%) |
|------------|---------------|-------------------------|
| 5.96       | 39.72         | 39.72                   |
| 3.30       | 21.99         | 61.71                   |
| 2.55       | 17.01         | 78.72                   |
| 0.94       | 6.26          | 84.98                   |
| 0.72       | 4.80          | 89.78                   |
| 0.46       | 3.07          | 92.85                   |
| 0.38       | 2.53          | 95.38                   |
| 0.22       | 1.46          | 96.84                   |
| 0.18       | 1.23          | 98.07                   |
| 0.11       | 0.76          | 98.83                   |
| 0.09       | 0.60          | 99.43                   |
| 0.06       | 0.41          | 99.84                   |
| 0.01       | 0.09          | 99.93                   |
| 0.01       | 0.05          | 99.98                   |
| 0.00       | 0.02          | 100.00                  |
biomass, acid detergent fibre and neutral detergent fibre; and the minimum standard genetic values were used for the traits in vitro digestibility of the dry biomass, cellulose/lignin ratio, nitrogen level, and ash content. The other ideotypes (ideotypes 2, 3, to 8) were designed using the same thinking, considering desirable and undesirable combining factors, as shown in Table 3.

FAI-BLUP index

The FAI-BLUP index allows the genotypes ranking (associated with spatial probability) based on multitrait, free from multicollinearity. In addition, genetic correlation - positive or negative - within each factor was taken into account, preserving the relationships between traits, and giving biological meaning to the ideotypes.

In the elephant grass case, the ideotypes designed may indicate the selection of genotypes for different purposes, for example ideotype 1 is suitable for conversion to bioenergy via combustion, ideotype 2 is suitable for the production of second generation ethanol, and ideotypes 6 and 8 are suitable for forage uses (cut-and-carry and grazing, respectively), as presented and discussed by Rocha et al. (2017b).

A breeding programme based on ideotype focuses on multitrait simultaneously. This method differs from other multivariate approaches in plant breeding, such as Smith-Hazel classical index, which tends to focus on few traits and not on the underlying plant morphology and physiology. Focusing directly on few traits statistically simplifies the problem (Van Oijen & Höglund, 2016); however, important information may not be considered on the data analyses.

Simultaneous selection methods, such as the Smith-Hazel classical index and other indices derivative from this index may be used, as there is no multicollinearity problem in the phenotypic covariance matrix. When comparing the Smith-Hazel classical index with the FAI-BLUP index, the diagnosis of multicollinearity was carried out, and some traits of the Smith-Hazel classical

### Table 2

| Traits* | Factor | Qualitymetric | Phytometric | Energymetric | Communality |
|---------|--------|---------------|-------------|--------------|-------------|
| CEL     | -0.8261| -0.3126       | -0.2235     | 0.8301       |
| LIG     | -0.8907| 0.1204        | 0.2877      | 0.8906       |
| C/L     | 0.7122 | -0.2122       | -0.4309     | 0.7380       |
| ADF     | -0.9495| -0.1808       | -0.0733     | 0.9396       |
| NDF     | -0.8587| -0.2083       | -0.2666     | 0.8519       |
| IDBD    | 0.8346 | -0.1061       | -0.1749     | 0.7384       |
| N       | 0.5787 | 0.1529        | -0.1949     | 0.3960       |
| HGT     | -0.2027| -0.6747       | 0.5247      | 0.7716       |
| PHV     | -0.0728| -0.9386       | 0.0873      | 0.8940       |
| TGB     | -0.0115| -0.9240       | 0.0803      | 0.8604       |
| TDB     | -0.0805| -0.9282       | -0.1129     | 0.8808       |
| STD     | 0.0406 | -0.4377       | 0.7490      | 0.7543       |
| DM      | -0.2439| -0.1787       | -0.7777     | 0.6962       |
| ASH     | 0.4842 | 0.2058        | -0.6784     | 0.7371       |
| CAV     | -0.5259| -0.0282       | 0.7430      | 0.8295       |

Mean | – | – | – | 0.7872

*CEL = Cellulose (g Kg\(^{-1}\)); LIG = Lignin (g Kg\(^{-1}\)); C/L = Cellulose/Lignin ratio; ADF = Acid detergent fibre (g Kg\(^{-1}\)); NDF = Neutral detergent fibre (g Kg\(^{-1}\)); IDBD = in vitro digestibility of the dry biomass (g Kg\(^{-1}\)); N = Nitrogen (g Kg\(^{-1}\)); HGT = mean height (m); PHV = Phenotypic vigor (1–5); TGB = Total green biomass (Mg ha\(^{-1}\)); TDB = Total dry biomass (Mg ha\(^{-1}\)); STD = Stalk diameter (mm); DM = Dry matter (g Kg\(^{-1}\)); ASH = Ash (g Kg\(^{-1}\)) and CAV = Calorific value (MJ kg\(^{-1}\)).

### Table 3

| Ideotypes | First | Second | Third |
|-----------|-------|--------|-------|
| 1         | Desirable | Desirable | Desirable |
| 2         | Desirable | Desirable | Undesirable |
| 3         | Desirable | Undesirable | Desirable |
| 4         | Desirable | Undesirable | Undesirable |
| 5         | Undesirable | Desirable | Desirable |
| 6         | Undesirable | Desirable | Undesirable |
| 7         | Undesirable | Undesirable | Desirable |
| 8         | Undesirable | Undesirable | Undesirable |

© 2017 The Authors. GCB Bioenergy Published by John Wiley & Sons Ltd, 10, 52–60
index were discarded to solve this problem: total green biomass, cellulose/lignin ratio, neutral detergent fibre, and acid detergent fibre.

The multicollinearity among traits will certainly appear when handling several traits, providing inflated errors. Inflated errors result in inaccurate tests of significance for the predictors, meaning that important predictors may not be significant, even if they are truly influential, because if a predictor, rather than other collinear predictor is dropped from the model, the selection process may proceed on a wrong trajectory with the risk of erroneous conclusions (Dormann et al., 2013; Prunier et al., 2015).

Methods for detecting and solving multicollinearity problems established for multiple regression can also be applied in SEM (Lei & Wu, 2007). Thus, using the factor analysis, which produces uncorrelated or orthogonal axes among final factors, scores free from multicollinearity are generated (Dormann et al., 2013). In addition, there is no need to assign weights to FAI-BLUP index, as it occurs in the Smith-Hazel classical index and its derived indices. Therefore, the index proposed in this study solves the major problems of the Smith-Hazel classical index.

**FAI-BLUP index vs. SH-BLUP comparison**

Table 4 presents the comparison - via predicted genetic gain - between the indices. Predicted genetic gain for the qualitymetric traits factor considering the FAI-BLUP index allowed obtaining gains in desirable sense for all traits, and showed gains of greater magnitude when compared with the SH-BLUP index. The SH-BLUP index predicted genetic gains in undesirable sense for lignin, cellulose/lignin ratio, in vitro digestibility of the dry biomass, and nitrogen level.

The predicted gains for the phytometric traits factor are given in desired sense for all traits in both indices.

| Table 4 | Predicted genetic gain based on direct selection by genetic values, the SH-BLUP index and the FAI-BLUP index. |
|---------|----------------------------------------------------------------------------------------------------------------|
| **Factor** | **Traits** | **Genetic value** | **GCV(%)†** | **Predicted genetic gain (%)** | **Direct Selection** | **SH-BLUP** | **FAI-BLUP** |
| Qualitymetric | CEL | 41.23 | 1.85 | 1.81 | 0.56 | 1.18 |
| | LIG | 9.83 | 5.54 | 6.29 | -0.25 | 1.93 |
| | C/L | 4.26 | 4.93‡ | -5.66 | 0.81 | -0.88 |
| | ADF | 51.58 | 2.04‡ | 1.97 | 0.42 | 1.22 |
| | NDF | 79.08 | 1.26‡ | 1.27 | 0.18 | 0.71 |
| | IDBD | 33.28 | 3.17 | -11.08 | 8.12 | -1.38 |
| | N | 0.45 | 8.55 | -7.83 | 1.76 | -3.77 |
| Phytometric | HGT | 3.56 | 8.36 | 10.93 | 4.24 | 7.19 |
| | PHV | 3.33 | 9.21 | 11.64 | 8.12 | 8.91 |
| | TGB | 54.93 | 24.81† | 34.24 | 31.52 | 24.46 |
| | TDB | 20.75 | 25.57 | 35.58 | 32.89 | 25.43 |
| Energy | STD | 16.90 | 13.44 | 18.26 | 7.14 | 8.59 |
| | DM | 386.00 | 6.41 | 6.88 | 0.72 | 0.37 |
| | ASH | 6.60 | 7.82 | -6.9 | -0.52 | -4.25 |
| | CAV | 17.66 | 0.59 | 0.61 | 0.02 | 0.31 |
| **Total desirable (positive) gain (%)§** | – | 66.08 | 62.01 |
| **Total desirable (negative) gain (%)¶** | – | -32.31 | 32.65 |

Selected genotype by SH-index- 95,2,55,21,53,51,84,70,93,1,40,72,44,96,26
Selected genotype by FAI-BLUP index - 2,95,55,96,6,78,60,53,39,21,88,50,3,41,70

*CEL = Cellulose (g Kg⁻¹); LIG = Lignin (g Kg⁻¹); C/L = Cellulose/Lignin ratio; ADF = Acid detergent fibre (g Kg⁻¹); NDF = Neutral detergent fibre (g Kg⁻¹); IDBD = In vitro digestibility of the dry biomass(g Kg⁻¹); N = Nitrogen (g Kg⁻¹); HGT = mean height (m); PHV = Phenotypic vigor (1–5); TGB = Total green biomass (Mg ha⁻¹); TDB = Total dry biomass (Mg ha⁻¹); STD = Stalk diameter (mm); DM = Dry matter (g Kg⁻¹); ASH = Ash (g Kg⁻¹) and CAV = Calorific value (MJ kg⁻¹).

†GCV = coefficient of genetic variation (%).
‡Traits excluded in the SH-BLUP, due to multicollinearity problems.
§Proportion of the predicted gain for traits which were desired to be increased (CEL, LIG, ADF, NDF, HGT, PHV, TGB, TDB, STD, DB and CAV) in relation to predicted gain for direct selection.
¶Proportion of the predicted gain for traits which were desired to be decreased (C/L, IDBD, N and ASH) in relation to predicted gain for direct selection.

© 2017 The Authors. GCB Bioenergy Published by John Wiley & Sons Ltd, 10, 52–60
The SH-BLUP index predicted greatest gains at once for total green biomass and total dry biomass (Table 4). This fact is explained due to greater relative economic weight attributed to the total dry biomass (Table 4) and to the high genetic correlation (positive) between total green biomass and total dry biomass (Table 2). However, the FAI-BLUP index estimated the greatest genetic gain for other traits.

The energymetric factor showed superior predicted genetic gain considering the FAI-BLUP index for stalk diameter, ash content, and calorific value. The FAI-BLUP index predicted lesser gain only for dry matter, when compared with the SH-BLUP index; however, all the gains were in the desired sense for both indices (Table 4).

Falconer & Mackay (1996) assigned two main causes of genetic correlations: pleiotropy and genetic linkage. The second is most appropriate to explain the predicted genetic gains using the FAI-BLUP index due to the quantitative nature of the traits and due to the factor analysis. According to Wang et al. (2009) the factor analysis explains the covariances or correlations among the traits in the factors; moreover, it can be used to understand what constructs underlie the data. Genetic linkage can only be broken by means of repeated cycles of meiosis (Luby & Shaw, 2009), and therefore desirable gains for all traits can be achieved considering the genotypes selected by the SH-BLUP index.

Direct selection provides the maximum predicted gain considering one trait at a time. In addition, the FAI-BLUP index provided gains of approximately 62% of direct selection, simultaneously for both traits that are desired to be increased, and of approximately 33% for traits that are desired to be decreased. The genotypes selected by the FAI-BLUP index have potential to improve all traits simultaneously, while the Smith-Hazel classical index predicted 66% and −32% for the traits that are desired to be increased and decreased, respectively, and it does not have potential to improve all traits simultaneously. The FAI-BLUP index provides more balanced genetic gains prediction for all traits.

The FAI-BLUP and the SH-BLUP indices share only seven of the 15 elephant grass with the highest performance for combustion (Table 4). Besides, the FAI-BLUP index without multicollinearity (i.e. when the traits that cause multicollinearity on the matrix of phenotypic variance-covariance are excluded) still provides superior gains in desirable sense for all traits, when compared with the SH-BLUP index. Thus, traits that cause multicollinearity work as auxiliary traits, improving the FAI-BLUP index.

**Elephant grass genotypes for energy cogeneration**

Elephant grass is a suitable crop to validate the multi-trait index proposal due to its multiple end uses (i.e. direct combustion, second generation ethanol production, grazing, cut-and-carry, silage, etc.). In order to select high performance genotypes for conversion to
bioenergy via combustion, elephant grass plants should bring forward high levels of: cellulose, lignin, acid detergent fibre, and neutral detergent fibre (McKendry, 2002; Casler & Jung, 2006; Dorez et al., 2014; Silveira et al., 2015), calorific value, dry matter (McKendry, 2002; Jaradat, 2010; Shao et al., 2012; Williams et al., 2016), mean height, phenotypic vigor, total green biomass, total dry biomass (Ward et al., 2008; Silveira et al., 2015), and stalk diameter (Rocha et al., 2017a); and low levels of: cellulose/lignin ratio, in vitro digestibility of the dry biomass, ash (McKendry, 2002; Casler & Jung, 2006; Dorez et al., 2014), and nitrogen (McKendry, 2002; Morris et al., 2012; Na et al., 2016).

Figure 1 shows the ranking of the 100 genotypes according to the FAI-BLUP index and its associated spatial probability. The results allowed single and easily selection process of genotypes. The first 15 genotypes selected according to the FAI-BLUP index (genotypes: 2, 95, 55, 96, 6, 78, 60, 53, 39, 21, 88, 50, 3, 41 and 70 – genotypes registrations see Table S1) present balanced traits related to conversion to bioenergy via combustion, and therefore should be selected as multitrait high performance genotypes.

In this study, a multitrait index (FAI-BLUP index) is proposed based on the structural equation models theory. The FAI-BLUP index indicated the 15 elephant grass genotypes with the highest performance for combustion based on multitrait, without assigning weights, free from multicollinearity, and balanced genetic gains were predicted in desirable sense for all measured traits. The Smith-Hazel classical index predicted genetic gains in undesirable sense for lignin, cellulose/lignin ratio, in vitro digestibility of the dry biomass and nitrogen. Moreover, in general, the genetic gains predicted by the FAI-BLUP index were superior to those predicted by the Smith-Hazel classical index. Therefore, the FAI-BLUP index is a new technical advance tool, and it can be applied in genetic breeding programmes. In addition, a routine analysis applied to R free software has been elaborated, so the breeder can use the FAI-BLUP index on his own data.

Acknowledgements
The authors thank Tiago de Souza Marçal for contributing in the FAI-BLUP index routine for the R software, and the agencies CAPES, FAPEMIG and UNIPASTO for the financial support.

References
AOAC. Association of Official Analytical Chemical (1975) Official methods of analysis of the Association of Official Analytical Chemists (12 edn). AOAC Association of Official Analytical Chemical, Washington, D.C.
Bartlett MS (1938) Methods of estimating mental factors. Nature, 141, 609–610.
Blering LL, Laviola BG, Salgado CC, Sanchez CFB, Rosado TT, Alves AA (2012) Genetic gains in physic nut using selection indexes. Pesquisa Agropecuaria Brasileira, 47, 402–408.
Casler MD, Jung HIG (2006) Relationships of fibre, lignin, and phenolics to in vitro fibre digestibility in three perennial grasses. Animal Feed Science and Technology, 125, 151–161.
Cerin-Rojas JJ, Crossa J, Sahagín-Castellanos J, Castillo-González F, Santacruz-Varela A (2006) A selection index method based on eigenanalysis. Crop Science, 46, 1711–1721.
Cruz CD (2013) GENES - a software package for analysis in experimental statistics and quantitative genetics. Acta Scientiarum. Agronomy, 5, 271–276.
Dorez G, Ferry L, Sonnier R, Taguet A, Lopez-Cuesta JM (2014) Effect of cellulose, hemicellulose and lignin contents on pyrolysis and combustion of natural fibers. Journal of Analytical and Applied Pyrolysis, 107, 325–331.
Dormann CF, Elith J, Bacher S et al. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36, 27–46.
Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman Group Limited, Edinburgh.
Fontoura CF, Brandão LE, Comes LL (2015) Elephant grass biorefineries: towards a cleaner Brazilian energy matrix? 2015. Journal of Cleaner Production, 86, 85–93.
Gianola D, Sorensen D (2004) Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. Genetics, 167, 1407–1424.
Goering HK, Van Soest PJ (1967) Forage fiber analysis: apparatus, reagents, procedures and some applications. Agricultural Handbook, Washington, D.C.
Hazel LN (1943) The genetic basis for constructing selection indexes. Genetics, 28, 476–490.
Hazel LN, Dickerson GE, Freeman AE (1994) The selection index - Then, now, and for the future. Journal of Dairy Science, 77, 3236–3251.
Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics, 31, 423–447.
Jaradat AA (2010) Genetic resources of energy crops: biological systems to combat climate change. Australian Journal of Crop Science, 4, 309–323.
Kaiser H (1958) The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23, 187–200.
Lei PW, Wu Q (2007) An NCME instructional module on introduction to structural equation modeling: issues and practical considerations. Educational Measurement: Issues and Practice, 26, 33–44.
Luby J, Shaw DV (2009) Plant breeders’ perspectives on improving yield and quality traits in horticultural food crops. HortScience, 44, 20–22.
McKendry P (2002) Energy production from biomass (part 1): overview of biomass. Bioresource Technology, 83, 37–46.
Montgomery DC, Peck EA (1992) Introduction to linear regression analysis. John Wiley and Sons, New York.
Morais RF, Quesada DM, Reis VM, Uruquiaga S, Alves BJR, Boddie RM (2012) Contribution of biological nitrogen fixation to Elephant grass (Pennisetum purpureum Schum.). Plant and Soil, 356, 23–34.
Na C, Sellenberger LE, Fedenko JR, Erickson JE, Woodard KR (2016) Seasonal changes in chemical composition and leaf proportion of elephantgrass and energycane biomass. Industrial Crops and Products, 94, 107–116.
Patterson HD, Thompson R (1971) Recovery of inter-block information when blocks sizes are unequal. Biometrika, 58, 454–455.
Prunier JG, Colyn M, Legendre X, Nimon KF, Flamand MC (2015) Multicollinearity in spatial genetics: separating the wheat from the chaff using commonality analysis. Molecular Ecology, 24, 263–263.
R Development Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org (accessed 7 October 2016).
Ra K, Shiotsu F, Abe J, Morita S (2012) Biomass yield and nitrogen use efficiency of cellulosic energy crops for ethanol production. Biomass & Bioenergy, 37, 330–334.
Rengsirikul K, Ishii Y, Kangsvaaschol K et al. (2013) Biomass yield, chemical composition, and potential ethanol yields of 8 cultivars of Napiergrass (Pennisetum purpureum Schumach.) harvested 3-monthly in central Thailand. Journal of Sustainable Bioenergy Systems, 3, 107–112.
Resende MDV (2007) Selegen-REML/BLUP: sistema estatístico e seleção genética computadorizado via modelos lineares mistos. Embrapa Florestas, Colombo.
Rocha JRASC, Machado JC, Carneiro PCS, Carneiro JC, Resende MDV, Pereira AV, Carneiro JES (2017a) Elephant grass ecotypes for bioenergy production via direct combustion of biomass. Industrial Crops and Products, 95, 27–32.
Rocha JRASC, Machado JC, Carneiro PCS, Carneiro JC, Resende MDV, Ledo FJS, Carneiro JES (2017b) Bioenergetic potential and genetic diversity of elephantgrass viamorpho-agaromonic and biomass traits. Industrial Crops and Products, 95, 485–492.
Rosa GJM, Valente BD, de los Campos G, Wu XL, Gianola DS, Silva MA (2011) Inference causal phenotype networks using structural equation models. Genetics Selection Evolution, 43, 1–13.
Santchurn D, Ramdoyal K, Badaloo MGH, Labuschagne M (2012) From sugar industry to cane industry: investigations on multivariate data analysis techniques in the identification of different high biomass sugarcane varieties. *Euphytica*, 185, 543–558.

Santchurn D, Ramdoyal K, Badaloo MGH, Labuschagne M (2014) From sugar industry to cane industry: evaluation and simultaneous selection of different types of high biomass canes. *Biomass & Bioenergy*, 61, 82–92.

Shao Y, Wang J, Preto F, Zhu J, Xu CC (2012) Ash deposition in biomass combustion or co-firing for power/heat generation. *Energies*, 5, 5171–5189.

Silva DJ, Queiroz AC (2012) *Aulas de alimentos (métodos químicos e biológicos)*, (3rd edn). Editora UFV, Vicsosa, MG.

Silveira LCI, Brasilheiro BP, Kist V, Weber H, Daros E, Peternelli LA, Barbosa MHP (2015) Selection strategy in families of energy cane based on biomass production and quality traits. *Euphytica*, 204, 443–455.

Smith HF (1936) A discriminant function for plant selection. *Annals of Eugenics*, 7, 240–250.

Stephens MJ, Alsopch PA, Beatson RA, Winfield C, Buck EJ (2012) Genetic parameters and development of a selection index for breeding red raspberries for processing. *Journal of the American Society of Horticultural Science*, 137, 236–242.

Tilley JHA, Terry RA (1963) A two stage technique for in vitro digestion of forage crops. *Journal of the British Grassland Society*, 18, 104–111.

Ullman JB (2006) Structural equation modeling: reviewing the basics and moving forward. *Journal of Personality Assessment*, 87, 35–50.

Valente BD, Rosa GJM, de los Campos G, Gianola D, Silva MA (2010) Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*, 185, 633–644.

Valente BD, Rosa GJM, Gianola D, Wu XL, Weigel K (2013) Is structural equation modeling advantageous for the genetic improvement of multiple traits? *Genetics*, 194, 561–572.

Van Oijen M, Hogland M (2016) Toward a Bayesian procedure for using process-based models in plant breeding, with application to ideotype design. *Euphytica*, 207, 627–643.

Wang X, Kammerer CM, Anderson S, Lu J, Feingold E (2009) A comparison of principle component analysis and factor analysis strategies for uncovering pleiotropic factors. *Genetic Epidemiology*, 33, 325–331.

Ward SM, Gaskin JF, Wilson LM (2008) Ecological genetics of plant invasion: what do we know? *Invasive Plant Science and Management*, 1, 98–109.

Williams CL, Westover TL, Emerson RM, Tumuluru JS, Li C (2016) Sources of biomass feedstock variability and the potential impact on biofuels production. *Bioenergy Research*, 9, 1–14.

Zhang W, Xu H, Zhu J (2009) Index selection on seed traits under direct, cytoplasmic and maternal effects in multiple environments. *Journal of Genetics and Genomics*, 36, 41–49.

**Supporting Information**

Additional Supporting Information may be found online in the supporting information tab for this article:

- **Table S1.** Registrations names of genotypes of the Active Elephantgrass Germplasm Bank (BAGCE) maintained by Embrapa Dairy Cattle Research Center and their respective code.

- **Routine S1.** FAI-BLUP index routine analysis applied to R software.