American Journal of Plant Sciences, 2018, 9, 216-230
http://www.scirp.org/journal/ajps
ISSN Online: 2158-2750
ISSN Print: 2158-2742

Decision Trees as a Tool to Select Sugarcane Families

Luiz A. Peternelli[1*], Diego P. Bernardes[1], Bruno P. Brasileiro[2], Marcio H. P. Barbosa[3], Raphael H. T. Silva[1]

[1]Department of Statistics, Universidade Federal de Vicosa, Vicosa, Brazil
[2]Department of Crop Science and Phytosanitary, Universidade Federal do Paraná, Curitiba, Brazil
[3]Department of Crop Science, Universidade Federal de Vicosa, Vicosa, Brazil
Email: *peternelli@ufv.br

How to cite this paper: Peternelli, L.A., Bernardes, D.P., Brasileiro, B.P., Barbosa, M.H.P. and Silva, R.H.T. (2018) Decision Trees as a Tool to Select Sugarcane Families. *American Journal of Plant Sciences*, **9**, 216-230.
https://doi.org/10.4236/ajps.2018.92018

Received: October 1, 2017
Accepted: January 22, 2018
Published: January 25, 2018

Abstract

New strategies are required in the sugarcane selection process to optimize the genetic gains in breeding programs. Conventional selection strategies have the disadvantage of requiring the weighing of all the plants in a plot or a sample of stalks and the counting of the number of stalks in all the experimental plots, which cannot always be performed because more than 200,000 genotypes routinely comprise the first test phase (T1) of most sugarcane breeding programs. One way to circumvent this problem is to use decision trees to rank the yield components (the stalk height, the stalk diameter and the number of stalks) and to subsequently use this categorization to select the best families for a specific trait. The objective of this study was to evaluate the categorization of yield components using the classification and regression tree (CART) algorithm as a family selection strategy by comparing the performance of CART with those of conventional methods that require the weighing of stalks, such as the best linear unbiased prediction (BLUP) with sequential (BLUPS) or individual simulated (BLUPIS) procedures. Data from five experiments performed in May 2007 in a randomized block design were analyzed. Each experiment consisted of five blocks, 22 families and two controls (commercial varieties). CART effectively defined the classes of the yield components and selected the best families with an accuracy of 74% compared to BLUPS and BLUPIS. Families with at least 11 stalks per linear meter of furrow resulted in productivities that were above the average productivity of the commercial varieties used in this study and are, therefore, recommended for selection.

Keywords

Statistical Learning, Plant Breeding, Saccharum Spp., Synthetic Data, Supervised Learning

1. Introduction

Genetic breeding programs are central to the sugarcane agribusiness. The use of novel cultivars can increase the average productivity of the Brazilian sugar and alcohol sector and improve the quality of the raw materials used in the production of sugar and ethanol [1].

Sugarcane genetic breeding programs usually consist of three test phases (T1, T2 and T3), an experimental phase (EP) and a multiplication phase [2]. Briefly, the first plant selections are performed in the T1 phase. A clone is selected in this phase that is cultivated in the subsequent phases through vegetative propagation. The clones are planted in experimental designs with replicates to identify potentially superior clones. After 8 to 10 years of evaluation, the best clones are used in final evaluation experiments (EP) in different locations, wherein the clones are evaluated for 3 to 5 harvests.

Although individual visual selection is routinely applied in the early phases of breeding programs [1] [3], this type of selection has been criticized for its inefficiency in terms of the absence of replicates, plant competition and confounding environmental effects [4]. The aforementioned authors have advocated the use of family selection followed by individual selection to produce greater gains than that obtained via mass selection, especially for low-heritability traits.

Along this line of thought, some breeding programs have prioritized family selection followed by individual selection to find superior clones [3] [4] [5]. This strategy is motivated by the higher likelihood of finding individuals with favorable traits in families with high genotypic values [5].

Reference [6] has shown that predicting genotypic values using the best linear unbiased predictor (BLUP) at individual level (BLUPI) procedure is the optimal sugarcane selection strategy. This procedure simultaneously uses information from families and individuals within families for selection. However, this procedure is seldom used in breeding programs because of the difficulty of collecting data on an individual level.

Some strategies to overcome this practical problem have been reported in the literature. Reference [3] developed what we shall call sequential BLUP (BLUPS). Families are ranked according to the trait being evaluated (usually tons of stalks per hectare—TSH), and the selection is performed for 40% of the families. The families comprising the 40% with the highest mean TSH are split into four groups. In the group of families with the highest means, 40% of the individuals from each family are selected, and in the other three groups, 30%, 20% and 10% of individuals are selected from each family. Reference [7] proposed the selection of families with genotypic values greater than the overall mean, followed by the simulation of the number of individuals to be selected in each family according to the ratio between the genotypic values of the families and the number of individuals to be selected in the best family. The latter procedure is termed BLUP individual simulated (BLUPIS).

The difficulties encountered in using BLUPS [3] and BLUPIS [7] in in-

ter-family and intra-family selection are related to the large volume of data that must be collected and the logistics that are required for timely data collection and processing to perform the selection because the data are collected at the end of the crop cycle. At least one representative sample of stalks from each experimental plot must be weighed to use these methods. The difficulties in finding skilled labor and operating costs often restrict the number of families that can be evaluated in the field.

Alternative data collection methods have been sought to streamline the family selection process by circumventing having to weigh plants from all the plots. Thus, a definition of classes (categorization) for the variables that incorporates crop yield components (the number of stalks, the stalk diameter and the stalk height) could significantly reduce the time expended on data collection, if such a definition were properly defined and experimentally validated. Decision trees can be used to categorize the yield components, specifically by using the classification and regression trees (CART) algorithm [4], which is a statistical method potentially useful for identifying families with the highest yield potential by combining classes of variables.

CART involves non-parametric statistical methods that are used in data partitioning through specific rules performed by binary divisions [8]. The objective of this technique is to describe the variability in the dependent variable as a function of the independent variables through binary divisions [9]. Reference [10] has argued that the advantage offered by this technique is that the algorithm evaluates all the possible predictors and divisions. Furthermore, the algorithm may be applied to other data sets that include the same variables used in designing the decision tree.

The objective of this study was to examine the efficiency of categorizing sugarcane yield components using the CART algorithm for sugarcane family selection to further the development of alternative data collection methods and reduce costs in the initial phase (T1) of sugarcane breeding programs. The efficiency of the algorithm was measured by comparison with the selection performed using conventionally used procedures *i.e.*, BLUPS and BLUPIS.

2. Material and Methods

2.1. Plant Material

In 2006, 110 full-sib families were assessed from biparental crosses performed at the Serra do Ouro Experimental Station of the Federal University of Alagoas, located in the municipality of Murici, Alagoas, Brazil.

Following acclimatization, the seedlings resulting from the crosses were used in experiments on families in the experimental field of the Sugarcane Research and Breeding Center at the Federal University of Viçosa, located in the municipality of Oratórios, Minas Gerais, Brazil at a latitude of 20˚25'S, a longitude of 42˚48'W and a 494-m altitude in a LVe soil. Oratórios has a climate classified as Aw according to Köppen and Geiger. The annual average temperature and rain-

fall are respectively 21.6˚C and 1162 mm.

Five experiments were performed in May 2007 using a randomized complete block design. Each experiment consisted of five blocks, 22 families and two controls (commercial varieties). The same controls were used in all the experiments. Each plot consisted of 20 plants, which were distributed in two 5-m-long furrows, 1.40 m apart, totaling 12,000 plants. Each family was thus represented by 100 genotypes, which is considered to be a sufficient number for selection within the best families [11]. Agronomic practices including weed control and soil fertilization were the usual for this crop at the experimental station. Field was not irrigated.

2.2. Data Collection

In 2009, the mean stalk height (SH) and stalk diameter (SD) of all plots of the five experiments were assessed. Stalk height (SH) was measured in meters for one stalk from each clump in the plot from the base to the first visible dewlap. Stalk diameter (SD) was measured in centimeters using a digital caliper in the third internode from the stalk base to the apex of one stalk per clump in the plot. In addition to the stalk height and diameter, the total number of stalks per plot (NS) was also counted.

The total plot mass (*TPM*), in kg, was determined by weighing all the stalks using a dynamometer. The stalk productivity, in tons of sugarcane per hectare (*TSH*), was estimated using the equation $TSH = TPM \times 10/PA$, where TPM is the total mass of the plot in kg, and PA is the plot area in m^2. In the present study, $PA = 14\ m^2$.

2.3. Selection Using CART

In this study, regression trees were used to create classes for the three yield component variables. Only the SH, SD, NS and TSH data of the controls that were tested in the experiments, totaling 50 observations, were used in designing the regression trees. However, since regression trees may be incorrectly generated or, in an extreme case, even not generated, if the number of observations is too small, we decided to also simulate control data prior to using the CART algorithm, resulting in a procedure known as "data synthesis". The use of synthetic data to improve the amount of data for comparing statistical methods or techniques has been previously used in other research works [12] [13] [14].

The simulation was performed based on the covariance matrix $\Sigma(4 \times 4$, positive definite) of the variables TSH, NS, SH and SD of two of the controls that were used in all five experiments. In the simulation algorithm, the Cholesky decomposition of the covariance matrix Σ was used to generate $\Sigma = CC'$, where C is a lower triangular matrix $m \times m$ known as the Cholesky factor. A normal multivariate vector $X = \mu + CZ$ was simulated, where μ is the mean vector of the controls, C is the Cholesky factor derived from the covariance matrix Σ, and Z is a vector of random independent and identically distributed (IID) variables

with a standard normal distribution. This procedure was used to generate 1000 row vectors of the type $\left[X_{i1}, X_{i2}, X_{i3}, X_{i4}\right]$, wherein X_{ij} (i = 1 to 1000, and j = 1 to 4) represents the simulated value of the variable j (TSH, NS, SD or SH) for individual i. The algorithm presented ensured that these four variables had a covariance matrix Σ and a mean vector μ [15] [16] [17].

The generated data were subsequently subjected to the standard CART algorithm procedure. The NS distribution is discrete (Poisson) and is characterized by a parameter λ = mean number of stalks per plot; however, this distribution can be approximated by a normal distribution [18] because λ is relatively large (mean = 111.74). Thus, the simulated value was approximated to the nearest integer. Tree pruning was performed according to the 1-SE rule [8] and 10-fold cross-validation [19] methods to generate more accurate estimates, to reduce over fitting and to facilitate the interpretation of results. In summary, regression trees were obtained using simulated data based on the control observations (1000 observation vectors), and pruned (according to the 10-fold cross-validation and the 1-SE rule methods) and unpruned regression trees were obtained using non-simulated data (50 observation vectors).

Combinations of variables that could produce TSH levels higher than the mean productivity of both controls were located in the generated trees to obtain a clone selection cutoff point. The intra-family selection procedure was subsequently defined as follows: the selected families were split into three classes to define the number of individuals to be selected in each family, as indicated by the CART algorithm. The classes were defined based on the number of replicates (plots) in which the family was selected by the algorithm. The family was selected in each plot (replicate) when the combination of variables used in the classifier met the selection criterion defined by the designed regression tree. The first class consisted of the families selected by CART in all five plots (or replicates) of the family. The second class consisted of the families selected in four replicates. Finally, the third class consisted of the families selected in three replicates. Thus, for the intra-family selection, 30% of the individuals from each family were selected in the best class, followed by 20% and 10% of the individuals from each family in the second and third classes, respectively. Note that other ratios could have been chosen, which could modify the results presented here. The choice reported herein was based on the aforementioned BLUPS procedure. In future studies, we will analyze the best selection ratio within our proposed use of CART.

2.4. Selection Using BLUPS and BLUPIS

The TSH data were analyzed using restricted maximum likelihood (REML)/BLUP mixed models and a statistical model associated with genotype assessment in an incomplete block design at plot means level by considering the matrix equation [6] $y = Xr + Zg + Wb + e$. In this equation y represents the data vector $\left(y \sim N\left(Xr, V\right)\right)$; r is the presumed fixed effects vector; g is the geno-

typic effects vector (presumed to be random), where $g \sim N(0, G)$ and G = the genetic covariance matrix of genotypes $\left(G = I\sigma_g^2\right)$; b is the environmental effects vector of the incomplete blocks (presumed to be random), where $b \sim N\left(0, I\sigma_b^2\right)$; and e is the vector of errors or residuals (random), where $e \sim N(0, R)$, R = residual covariance matrix $\left(R = I\sigma_e^2\right)$. X, Z and W are the incidence matrices for the said effect. The variance components σ_g^2, σ_b^2 and σ_e^2 correspond to the genotypic variance, the block variance and the residual variance, respectively.

The selection in the BLUPS procedure was performed following the strategy used by the Australian breeding program [3] to select 40% of the families tested. The selected families were split into four classes based on the TSH means. Each class consisted of 11 families, and 40% of the individuals within each family of the first class and 30%, 20% and 10% of individuals in each family in classes 2, 3 and 4 were selected, respectively.

In the BLUPIS procedure, the families with TSH means higher than the overall mean were selected [7]. The number of individuals selected from each family k (k = 1 to 52) was calculated using $n = \left(\hat{g}_k / \hat{g}_j\right) n_j$, wherein $\hat{g}_k$ refers to the estimated genotypic value of family k, $\hat{g}_j$ refers to the estimated genotypic value of the best family, and n_j is the number of individuals selected from the best family. In the present study, n_j = 27 individuals were selected from the best family. A mixed models analysis was performed using the SELEGEN-REML/BLUP software [20].

2.5. Comparison between BLUPS, BLUPIS and CART

Confusion matrices were generated for each tree to facilitate the visualization of the similarities and differences among the selection methods BLUPS and BLUPIS (which were considered as conventional methods and, therefore, considered correct and were subsequently used for comparison purposes) and CART (the method being tested) (**Figure 1**).

This confusion matrix was used to calculate four useful statistical parameters to assess the applicability of the selection method: 1) the choice accuracy (CAc), where CAc = (A + D)/T_{ABCD}; 2) the apparent error rate (AER), where AER = 1 − CAc; 3) the selection precision (SeP), where SeP = A/T_{AC}; and 4) the error of omission (EOm), where EOm = 1 − SeP.

		Conventional method		
		Selects	Fails to select	Total
Tested method	Selects	A	B	T_{AB}
	Fails to select	C	D	T_{CD}
	Total	T_{AC}	T_{BD}	T_{ABCD}

Figure 1. Schematic of a confusion matrix showing frequencies of occurrence (A, B, C and D) for combinations of classes (Selects or Fails to select): the "conventional method" corresponds to the method used in practice, which is considered to be "ideal" or "true"; the "Tested method" corresponds to the novel method that was developed in this study.

The selection obtained using BLUPIS or BLUPS was considered to be the correct selection in the comparisons because these procedures are routinely used in breeding programs.

The CAc refers to the number of families selected or not selected by CART and BLUPS or BLUPIS relative to the total number of experimental families. The AER corresponds to the selection error. The SeP is the number of families simultaneously selected by CART, BLUPIS or BLUPS divided by the total number of families selected by BLUPS or BLUPIS. Finally, EOm is the error relative to the failure to select some families, as indicated by BLUPS or BLUPIS.

All the analyses and graphs of the CART algorithm were generated in the free software R [21] using the package rpart() [22].

3. Results and Discussion

Table 1 outlines the number of individuals selected from the families with the highest TSH means according to the BLUPS, BLUPIS and CART strategies using the original data (without simulation) and via CART after increasing the volume of control data through simulation. Whereas the families were ranked based on the TSH genotypic means obtained using the BLUPS and BLUPIS procedures, the families selected using CART were ranked based on the number of replicates in which each family was indicated for selection. A total of 52 families were selected using the BLUPIS procedure, corresponding to families that had genotypic means higher than the overall mean of the original population (102 $t \cdot ha^{-1}$). A total of 44 families were selected using the BLUPS procedure, corresponding to 40% of the 110 families considered in this study. CART selected 52 families when simulation was not used and 49 families following simulation (Table 1 and Table 2).

Although all the yield components (NS, SD and SH) were used to generate the regression trees, CART discarded the components SH and SD when predicting the TSH values. This result indicates that, according to the data analysis, the number of stalks was the variable that most strongly affected the productivity. Various studies on sugarcane path analysis and logistic regression have shown that NS is more important than other yield components [23] [24] [25]. The aforementioned authors have reported that families and clones with high TSH values can be successfully selected using NS only because NS is the main determinant of variation in TSH.

For the selection intensity used, BLUPS indicated 40 individuals in the best family for selection, whereas BLUPIS indicated 27 individuals, and CART indicated 30 individuals. Considering the intra-family selection criteria defined in this study, a total of 1100 individuals were indicated for selection by BLUPS, 1077 by BLUPIS, 1022 by CART using non-simulated data, and 890 by CART using simulated data (Table 1).

Table 2 shows the confusion matrices among CART, BLUPS and BLUPIS and the respective measures used to assess the CART efficiency. In the specific

Table 1. Genotypic TSH means (u + g) of families selected using BLUPS, BLUPIS and CART using data with and without simulation, number of replicates (plots) wherein each family was selected using CART (Rep) and number of individuals selected within each family (n_k).

Order	Without simulation						With simulation		
	Family	u + g	Rep	n_k			Family	Rep	n_k
				BLUPS	BLUPIS	CART			CART
1	28	157.0236	5	40	27	30	28	5	30
2	90	156.0278	5	40	27	30	90	4	20
3	42	151.0156	4	40	26	20	**42**	2	0
4	75	150.6422	3	40	26	10	75	3	10
5	69	150.0833	4	40	26	20	69	5	30
6	39	140.8769	5	40	24	30	39	5	30
7	113	139.3131	5	40	24	30	113	4	20
8	117	137.7323	5	40	24	30	117	5	30
9	106	137.2079	5	40	24	30	106	5	30
10	**70***	136.7379	1	40	24	0	**70**	1	0
11	38	133.924	5	40	23	30	38	5	30
12	26	130.6799	3	30	22	10	26	3	10
13	2	127.906	5	30	22	30	2	5	30
14	61	127.8761	5	30	22	30	61	5	30
15	78	127.8528	4	30	22	20	78	4	20
16	27	125.4197	5	30	22	.30	27	5	30
17	34	124.4823	4	30	21	20	34	4	20
18	66	124.1805	5	30	21	30	66	5	30
19	100	123.5837	3	30	21	10	100	3	10
20	**89**	121.6171	1	30	21	0	**89**	1	0
21	81	121.2884	3	30	21	10	81	3	10
22	12	120.4981	4	30	21	20	12	4	20
23	**29**	119.5196	1	20	21	0	**29**	1	0
24	**43**	118.4544	2	20	20	0	**43**	2	0
25	65	117.1188	4	20	20	20	65	4	20
26	67	116.4488	4	20	20	20	67	3	10
27	7	115.7334	3	20	20	10	7	3	10
28	**80**	115.419	1	20	20	0	**80**	1	0
29	71	114.8461	4	20	20	20	71	4	20
30	50	114.7471	5	20	20	30	50	5	30
31	25	114.2707	4	20	20	20	25	3	10
32	**23**	114.2551	2	20	20	0	**23**	2	0

Continued

33	54	114.1181	3	20	20	10	54	3	10
34	**84**	114.0369	1	10	20	0	**84**	2	0
35	88	113.5553	3	10	20	10	88	3	10
36	**47**	111.458	0	10	19	0	**47**	0	0
37	9	108.5093	4	10	19	20	9	3	10
38	111	108.037	3	10	19	10	111	3	10
39	76	107.3001	4	10	18	20	76	4	20
40	**94**	106.6959	2	10	18	0	**94**	2	0
41	35	105.795	3	10	18	10	**35**	1	0
42	96	105.5061	4	10	18	20	96	4	20
43	**63**	105.3483	2	10	18	0	**63**	0	0
44	53	105.1904	4	10	18	20	53	4	20
45	6	104.6934	3	0	18	10	6	3	10
46	**72**	104.3787	1	0	18	0	**72**	1	0
47	**68**	104.1732	2	0	18	0	**68**	2	0
48	14	104.0509	4	0	18	20	14	4	20
49	24	103.9342	3	0	18	10	24	3	10
50	**118**	103.8421	1	0	18	0	**118**	1	0
51	95	103.7601	4	0	18	20	95	4	20
52	**101**	103.5018	0	0	18	0	**101**	2	0
53	22	101.9428	5	0	0	30	22	4	20
55	1	101.2394	5	0	0	30	1	4	20
56	55	101.1084	4	0	0	20	55	3	10
60	56	99.605	4	0	0	20	56	3	10
61	20	99.4583	3	0	0	10	20	3	10
63	45	98.8018	4	0	0	20	45	4	20
65	103	97.8211	4	0	0	20	103	4	20
66	109	97.158	3	0	0	10	109	3	10
69	17	96.0743	4	0	0	20	17	3	10
71	64	95.8345	3	0	0	10	64	3	10
82	49	87.8862	4	0	0	20	49	4	20
83	4	85.0302	3	0	0	10	4	3	10
84	3	84.9651	4	0	0	20	3	4	20
93	11	75.5039	3	0	0	10	**11**	0	0
	Total			1100	1077	1020		**Total**	890

*Families not selected using CART because they failed to exhibit satisfactory results (≥11 stalks/m) in at least 50% of plots are shown in bold.

Table 2. Confusion matrices between the family selection strategies using CART, BLUPIS and BLUPS, together with measures of choice accuracy (CAc), apparent error rate (AER), selection precision (SeP) and error of omission (EOm) for the original data (without simulation) and following simulation (with simulation).

Without simulation						
CART	BLUPIS		Total	BLUPS		Total
	S	N		S	N	
S*	38	14	52	34	18	52
N	14	44	58	10	48	58
Total	52	58	110	44	66	110
CAc (AER)	0.745 (0.255)			0.745 (0.255)		
SeP (EOm)	0.731 (0.269)			0.773 (0.227)		
With simulation						
CART	BLUPIS		Total	BLUPS		Total
	S	N		S	N	
S	36	13	49	32	17	49
N	16	45	61	12	49	61
Total	52	58	110	44	66	110
CAc (AER)	0.736 (0.264)			0.736 (0.264)		
SeP (EOm)	0.692 (0.308)			0.727 (0.273)		

*S = selected families, N = non-selected families.

context of sugarcane family selection, the higher the choice accuracy (CAc) and the smaller the error of omission (EOm) are, the better is the CART performance. In a breeding program, the error of omission (EOm) is more compromising than the error of selecting more families improperly, that is, the error corresponding to B/TAB. The genotypes that pass to the next phase, coming from the families improperly selected by CART, would be subjected to new selection cycles within the breeding program, where these genotypes could then be excluded, if necessary. That is, the performance of CART improves for a higher number of correct predictions of selected and non-selected families (higher CAc), as indicated by the other procedures (BLUPS or BLUPIS), and a smaller number of families selected using BLUPS and BLUPIS and discarded by CART (smaller EOm).

Using non-simulated data, CART identified 38 of 52 families selected by BLUPIS (SeP = 0.731), that is, 73% of families with high TSH values (**Table 2**). CART failed to select 14 families selected by BLUPIS, resulting in an EOm = 0.269. Coincidentally, 14 other families not selected by BLUPIS were selected by CART. This error corresponds to another type of selection error, which is less compromising than the EOm because the genotypes selected in the respective families are assessed in the subsequent stages of the breeding program, where these genotypes may be eventually excluded from the breeding population, as

previously mentioned. Similar reasoning applies when comparing the apparent error from CART selection relative to BLUPS selection (EOm = 0.227, Table 2).

The CART choice accuracy values were similar to those obtained using BLUPIS and BLUPS (CAc = 0.745). In practical terms, this result indicates that CART successfully predicted 74.5% of the families selected or non-selected by BLUPS or BLUPIS, even when only using the number of stalks in the plot. This accuracy ratio is greater than 0.5 (p-value = 1.26e−07), a value that would be expected by chance if selection using CART had no relationship whatsoever with the other methods.

CART, based only on NS, indicated the selection of 52 families, 14 (26.9%) of which would not have been selected by BLUPIS and 18 of which would not have been selected by BLUPS (Table 2). When considering only potentially superior families, that is, those families that should be selected, CART exhibited significant selection precision compared to BLUPIS (SeP = 0.731, p-value = 0.0005976, H_0:π = 0.5) or BLUPS (SeP = 0.727, p-value = 0.0001941, H_0:π = 0.5). These percentages were relatively low but ensured that there was a reasonable amount of potential families in the subsequent stages of the breeding program at a rather reduced operation cost because only NS data were required.

According to [26], approximately 60% of the best genotypes are concentrated in 10% of the best families, and little can be gained by selecting more than 20% of the families. Therefore, the use of the CART algorithm and the selection rate from the BLUPIS and BLUPS methods should ensure the selection of 10% to 20% of the best families, and the best individuals would consequently be assessed in the second test phase (T2).

When considering only simulated data, the CART choice accuracy values were also similar to those obtained using BLUPIS or BLUPS, with CAc = 0.736. The results obtained using simulated data (synthetic data) were actually very similar to those obtained using non-simulated data, most likely because of the relatively large number of control data (a total of 25 plots per control, which contributed data for the CART algorithm). Using simulation data prior to the CART procedure has the potential advantage of enabling the means for ideotypes (ideal families) to be simulated at the researcher's discretion, which can be used to define which families to select from those present in a specific experiment. The results in Table 2 show the relevancy of the simulation procedure because the measures of the choice accuracy and the selection precision of the simulated and non-simulated data were rather similar, indicating sustained algorithm performance. Furthermore, the simulation can enable offsetting limited control data in a specific experiment. In the extreme case of the absence of controls, data could be simulated if the researcher is able to define a mean vector and a covariance matrix for the variables of interest according to the study population and considering the environment in which the selection is performed. This information could be retrieved from historical records from other experiments that have been conducted at the same location, for example, or from other studies reporting the information.

The use of tree pruning (1-SE rule or 10-fold cross-validation methods) to generate more accurate estimates resulted in no changes in the trees obtained, both for the simulated and non-simulated data. There was no change in the trees for which the pruning procedure was used because the algorithm could reach the optimal tree without using a fit to the model, which may have resulted from the good volume and quality of the data that were used in the analyses.

Figure 2 shows the regression tree with the non-simulated data generated by CART. The mean productivity of the controls was 145.81 TSH. Productivities higher than this value were generated by families with NS values higher than 110.5. That is, the NS was ranked into two classes, of which the first consisted of families with total NS values per plot below 110.5, and the second consisted of families with corresponding values above 110.5. This cutoff point between the classes corresponded to at least 11 stalks per linear meter of furrow because the plots consisted of two five-meter furrows.

Figure 3 shows the regression tree with the simulated data. The productivities generated for this tree were higher than 145.81 TSH when the total NS per plot was higher than 113.4, that is, at least 11 stalks per linear meter, which corroborated the result found using only the original data. However, the increase in the volume of data via simulation enabled additional classes of predicted values for TSH to be defined according to the total NS per plot of the family, which may be advantageous within the family selection process. Thus, it would be sufficient to select families with 13 to 15 stalks per linear meter if the breeder aims to select families with predicted TSH values ranging from 157 to 180 $t \cdot ha^{-1}$. A NS per linear meter above 15 and below 18 would indicate families with predicted TSH values ranging from approximately 180 to 200 $t \cdot ha^{-1}$. Families with more than 18 stalks per linear meter would be associated with predicted TSH values above 230 $t \cdot ha^{-1}$. Although the yield components SD and SH are not included in the regression tree generated by CART, the breeder should assess these traits and others, including the disease resistance, the lateral bud outgrowth, the internode length, the growth habits and other agronomic aspects of plants, for selection in families with higher productivity potential.

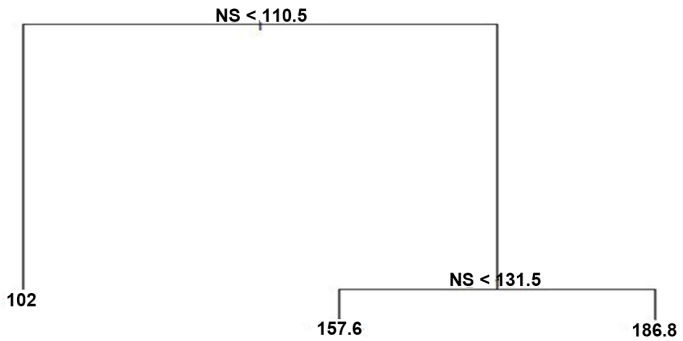

Figure 2. Regression trees generated using the CART algorithm for control data, wherein NS represents the total number of stalks per plot (two 5-m-long furrows), and the terminal nodes represent the predicted yield in tons of stalks per hectare (TSH); non-simulated data.

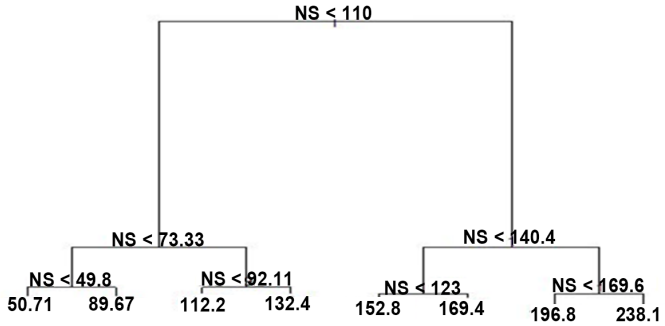

Figure 3. Regression trees generated using the CART algorithm for control data, wherein NS represents the total number of stalks per plot (two 5-m-long furrows), and the terminal nodes represent the predicted yield in tons of stalks per hectare (TSH); simulated data.

Table 3. Mean of the selected population (Ms), in tons of stalks per hectare (TSH), and number of families (nf) selected using the BLUPS, BLUPIS and CART selection strategies.

	BLUPS	BLUPIS	CART	
			NSim*	SimD
Ms	123.4621	120.4744	115.5319	115.8233
nf	44	52	52	49

*NSim = non-simulated data; SimD = simulated data.

The mean of the population selected by CART was lower than that selected by BLUP and BLUPIS for both the simulated and non-simulated data (Table 3). This result was obtained because CART selected families with TSH genotypic means below the overall mean of the original population (Table 1). However, the considerable advantage offered by CART is that the entire plot does not need to be weighed, which is necessary in the application of BLUPS or BLUPIS. Counting the number of stalks alone can be used to obtain a highly accurate selection of the best families when using CART.

The CART selection strategy may reduce operational costs because a smaller amount of manpower and a shorter execution time are required both to establish the experiments and to evaluate the families, which may result in a more efficient process of individual selection in the initial phases of sugarcane genetic breeding programs.

4. Conclusions

The CART algorithm effectively defined the classes of yield components followed by family selection with a mean accuracy of 74% compared to the BLUPIS and BLUPS selection procedures, which are usually applied in most sugarcane breeding programs.

A regression tree based only on the number of stalks per plot was sufficient to predict the sugarcane productivity classes. This study shows that families with more than 11 stalks per linear meter of furrow are potentially more productive

and should be selected and inspected for other agronomic characteristics.

Data simulation based on the covariance matrix between variables collected in controls had no effect on the results assessed in the present study because the NS showed a high correlation with the TSH.

Acknowledgements

We thank CNPq, FAPEMIG, and CAPES for financial support and RIDESA (Inter-University Network for the Development of Sugarcane Industry) and PMGCA-UFV, for providing the dataset.

References

[1] Barbosa, M.H.P., Resende, M.D.V., Dias, L.A.S., Barbosa, G.V.S., Oliveira, R.A., Peternelli, L.A. and Daros, E. (2012) Genetic Improvement of Sugar Cane for Bioenergy: The Brazilian Experience in Network Research with RIDESA. *Crop Breeding and Applied Biotechnology,* **12**, 87-98. https://doi.org/10.1590/S1984-70332012000500010

[2] Barbosa, M.H.P., Silveira, L.C.I. (2012) Breeding and Cultivar Recommendations. In: Santos, F., Borém, A. and Caldas, C. Eds., Sugarcane: Bioenergy, Sugar and Ethanol—Technology and Prospects. Suprema, Viçosa, MG, 568 p.

[3] Stringer, J.K., Cox, M.C., Atkin, F.C., Wei, X. and Hogarth, D.M. (2011) Family Selection Improves the Efficiency and Effectiveness of Selecting Original Seedlings and Parents. *Sugar Tech,* **13**, 36-41. https://doi.org/10.1007/s12355-011-0073-5

[4] Kimbeng, C.A. and Cox, M.C. (2003) Early Generation Selection of Sugarcane Families and Clones in Australia: A Review. *Journal American Society of Sugarcane Technologists*, **23**, 20-39.

[5] Resende, M.D.V. and Barbosa, M.H.P. (2005) Melhoramento genético de plantas de propagação assexuada. [Genetic Breeding of Plants with Asexual Propagation.] Embrapa, Colombo, Paraná.

[6] Resende, M.D.V. (2002) Genética biométrica e estatística no melhoramento de plantas perenes. [Biometric and Statistical Genetics in the Improvement of Perennial Plants.] Embrapa, Brasília.

[7] Resende, M.D.V. and Barbosa, M.H.P. (2006) Selection via Simulated Blup Based on Family Genotypic Effects in Sugarcane. *Pesquisa Agropecuária Brasileira,* **41**, 421-429. https://doi.org/10.1590/S0100-204X2006000300008

[8] Breiman, L., Friedman, J., Stone, C.J. and Olshen, S.A. (1984) Classification and Regression Trees. Chapman & Hall/CRC, Belmont, CA.

[9] Finch, H. and Schneider, M.K. (2007) Classification Accuracy of Neural Networks vs. Discriminant Analysis Logistic Regression, and Classification and Regression Trees. *Methodology*, **3**, 47-57. https://doi.org/10.1027/1614-2241.3.2.47

[10] Scholes D., Yu, O., Raebel, M.A., Trabert, B. and Holt, V.L. (2011) Improving Automated Case Finding for Ectopic Pregnancy Using a Classification Algorithm. *Human Reproduction*, **26**, 3163-3168. https://doi.org/10.1093/humrep/der299

[11] Leite, M.S.O., Peternelli, L.A., Barbosa, M.H.P., Cecon, P.R. and Cruz, C.D. (2009) Sample Size for Full-Sib Family Evaluation in Sugarcane. *Pesquisa Agropecuária Brasileira*, **44**, 1562-1574. https://doi.org/10.1590/S0100-204X2009001200002

[12] Moreira, E.F.A. and Peternelli, L.A. (2015) Sugarcane Families Selection in Early Stages Based on Classification by Discriminant Linear Analysis. *Revista Brasileira de*

Biometria, **33**, 484-493. https://doi.org/10.1590/S1984-70332013000200008

[13] Peternelli, L.A., Moreira, E.F.A., Nascimento, M. and Cruz, C.D. (2017) Artificial Neural Networks and Linear Discriminant Analysis in Early Selection among Sugarcane Families. *Crop Breeding and Applied Biotechnology*, **17**, 299-305. https://doi.org/10.1590/1984-70332017v17n4a46

[14] Nascimento, M., Peternelli, L.A., Cruz, C.D., Nascimento, A.C.C., Ferreira, R.P., Bhering, L.L., Salgado, C.C. (2013) Artificial Neural Network for Adaptability and Stability Evaluation in Alfalfa Genotypes. *Crop Breeding and Applied Biotechnology*, **13**, 152-156. https://doi.org/10.1590/S1984-70332013000200008

[15] Cressie, N.A.C. (2015) Statistics for Spatial Data. Revised Edition, John Wiley & Sons, Inc., New York. https://doi.org/10.1002/9781119115151

[16] Haining, R. (2005) Spatial Data Analysis—Theory and Practice. Cambridge University Press, Cambridge.

[17] Santos, A.C. and Ferreira, D.F. (2003) Sample Size Definition using Monte Carlo Simulation for the Normality Test Based on Skewness and Kurtosis. II. Multivariate Approach. *Ciência Agrotécnica*, **27**, 62-69. https://doi.org/10.1590/S1413-70542003000100007

[18] Casella, G. and Berger, R.L. (2002) Statistical Inference. 2nd Edition, Duxbury Press, Pacific Grove.

[19] Faraway, J.J. (2006) Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression. Chapman & Hall/CRC, New York.

[20] Resende, M.D.V. (2016) Selegen-REML/BLUP: A Useful Tool for Plant Breeding. *Crop Breeding and Applied Biotechnology*, **16**, 330-339. https://doi.org/10.1590/1984-70332016v16n4a49

[21] R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/

[22] Therneau, T., Atkinson, B. and Ripley, B. (2014) Rpart: Recursive Partitioning and Regression Trees. R Package Version 4.1-8. http://CRAN.R-project.org/package=rpart

[23] Brasileiro, B.P., Peternelli, L.A. and Barbosa, M.H.P. (2013) Consistency of the Results of Path Analysis among Sugarcane Experiments. *Crop Breeding and Applied Biotechnology*, **13**, 113-119. https://doi.org/10.1590/S1984-70332013000200003

[24] Espósito, D.P., Peternelli, L.A., Paula, T.O.M. and Barbosa, M.H.P. (2012) Path Analysis using Phenotypic and Genotypic Values for Yield Components in the Selection of Sugarcane Families. *Ciência Rural*, **42**, 38-44. https://doi.org/10.1590/S0103-84782011005000152

[25] Zhou, M.M., Kimbeng, C.A., Tew, T.L., Gravois, K.A., Pontif, M. and Bischoff, K.P. (2014) Logistic Regression Models to Aid Selection in Early Stages of Sugarcane Breeding. *Sugar Tech*, **16**, 150-156. https://doi.org/10.1007/s12355-013-0266-1

[26] Simmonds, N.W. (1996) Family Selection in Plant Breeding. *Euphytica*, **90**, 201-208. https://doi.org/10.1007/BF00023859