DIEGO PAIVA BERNARDES

SELEÇÃO DE FAMÍLIAS DE CANA-DE-AÇÚCAR VIA ÁRVORES DE DECISÃO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

VIÇOSA
MINAS GERAIS – BRASIL
2013

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
B522s
2013

Bernardes, Diego Paiva, 1985-
Seleção de famílias de cana-de-açúcar via árvores de decisão / Diego Paiva Bernardes. – Viçosa, MG, 2013.
ix, 29 f. : il. ; 29 cm.

Orientador: Luiz Alexandre Peternelli.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f . 25-29.

1. Saccharum – Seleção. 2. Estatística. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 633.61

DIEGO PAIVA BERNARDES

SELEÇÃO DE FAMÍLIAS DE CANA-DE-AÇÚCAR VIA ÁRVORES DE DECISÃO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

APROVADA: 15 de março de 2013

_____ _____

Volmir Kist

Márcio Henrique Pereira Barbosa
(Coorientador)

_____

Moysés Nascimento
(Presidente da banca)

À minha família

“A alegria está na luta, na tentativa, no sofrimento envolvido e não na vitoria propriamente dita.”

Gandhi

AGRADECIMENTOS

Agradeço a Deus por me dar saúde, sabedoria e paciência quando necessitei.

Aos meus pais, Gilson e Isis, meus irmãos, Samuel e Bruno, meu tio Putuca, por toda torcida, orgulho e apoio que me passaram a cada conquista.

Ao meu orientador, Luiz Alexandre Peternelli, por todas as oportunidades oferecidas desde a graduação.

A todos os amigos do mestrado, sem exceção.

Aos grandes amigos que fiz em Viçosa nesse período de pós-graduação, Vidigal, Elinho, Brasileiro e Rafael Rocha.

Vocês todos foram fundamentais nesta etapa.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

RESUMO

BERNARDES, Diego Paiva, M.Sc., Universidade Federal de Viçosa, março de 2013. **Seleção de famílias de cana-de-açucar via árvores de decisão.** Orientador: Luiz Alexandre Peternelli. Coorientadores: Márcio Henrique Pereira Barbosa e Moysés Nascimento.

O processo de seleção de clones de cana-de-açúcar é carente de métodos fitotécnicos e estatísticos que elevem os ganhos genéticos nos programas de melhoramento da cultura da cana-de-açúcar. De cinco fases de seleção do programa de melhoramento da RIDESA, a primeira é dotada de grande importância porque dela se desenvolvem as demais fases do programa e porque o volume de informações a ser analisados é enorme. Assim, caso os dados não forem corretamente analisados, pode-se descartar bons materiais logo nas primeiras fases, diminuindo a excelência do programa. As estratégias usuais de seleção, BLUP e BLUPIS, têm a desvantagem de exigir a de pesagem de toda a parcela. Uma maneira de se contornar isso é categorizar os componentes de produção; altura de colmos, diâmetro de colmos e número de colmos, via árvores de decisão. Através dessas árvores, é possível gerar as combinações desses componentes de produção e os respectivos valores de produção. Utilizando dados de testemunhas para gerar as árvores, não seria necessária a pesagem de toda a parcela, economizando tempo e recursos financeiros. O objetivo desse trabalho foi avaliar a categorização dos componentes de produção como estratégia de seleção entre e dentro de famílias através da comparação de seu desempenho com os métodos usuais, BLUP e BLUPIS. O algoritmo de árvore utilizado foi o CART. De natureza não paramétrica, esse é capaz de produzir divisões binárias combinando as variáveis explicativas e associando-as com distintos valores de resposta. Os dados foram coletados de 5 experimentos, instalados em maio de 2007, no delineamento em blocos casualizados, sendo cada experimento constituído de 5 blocos, 22 famílias e 2 testemunhas. O algoritmo CART foi eficiente em definir as classes dos componentes de produção seguido da seleção das melhores famílias no campo com acurácia média próxima de 73% quando comparado com o BLUPIS e BLUP.

ABSTRACT

BERNARDES, Diego Paiva, M.Sc., Universidade Federal de Viçosa, March, 2013. **Families selection through decision trees.** Adviser: Luiz Alexandre Peternelli. Co-Advisers: Márcio Henrique Pereira Barbosa and Moysés Nascimento.

The process of selection of clones to be used as new commercial varieties is lacking in statistical methods and phytotechnical that increase the genetic gains in crop improvement programs of cane sugar. Out of five stages of the selection program to improve the RIDESA, the first one, T1, is endowed with great importance. First because it develops the remaining phases of the crop breeding program. And second because the volume of information to be analyzed is huge. Thus, if the data is not properly analyzed, good clones can be ruled out in the early stages, reducing the excellence of the program. The common strategies of selection, BLUP and BLUPIS, have the disadvantage of the necessity of weighing the whole plot. One way around this is to categorize the components of production; stalk height, stalk diameter and number of stalks, by decision trees. Through these trees, you can generate the combinations of these yield components and their production values. Using data from commercial varieties to generate the trees, it would not be necessary to weigh the entire plot, saving time and money. The aim of this study was to evaluate the categorization of yield components as selection strategy between and within families by comparing their performance with the usual methods, BLUP and BLUPIS. The algorithm used was the CART. Non-parametric by nature, it`s capable of producing binary divisions combining the explanatory variables and associating them with different response values. Data were collected from 5 experiments, installed in May 2007 in a randomized block design, with each experiment consisting of 5 blocks, 22 families and two commercial varieties. CART algorithm was effective in defining classes of yield components followed by selection of the best families with mean accuracy of 73% when compared with BLUPIS and BLUP.

1. INTRODUÇÃO GERAL

O sistema agroindustrial da cana-de-açúcar é dotado de grande importância no cenário da economia nacional e mundial, não somente pela produção de açúcar, mas também pela demanda de fonte alternativa de energia para veículos automotores (NEVES; CONEJERO, 2007). Em 2010 a participação no PIB nacional do setor sucroalcooleiro foi de 1,56% (CONFEDERAÇÃO NACIONAL DAS INDÚSTRIAS, 2012). Isso se deve aos diversos produtos estratégicos originados desta cultura, como açúcar e etanol, representando uma fonte energética, e dos seus subprodutos, como o bagaço, a vinhaça ou o vinhoto, utilizados na produção de energia, alimentação animal e biofertilizantes.

A lavoura de cana-de-açúcar no Brasil se encontra em expansão, apresentando produtividade média de quase 69.000 kg/ha (COMPANHIA NACIONAL DO ABASTECIMENTO, 2012). O Estado de São Paulo lidera o ranking de produtores com 51,87% de área plantada, o que equivale a, aproximadamente, 4400 hectares. Logo atrás vem o Estado de Goiás com 8,52% de área plantada e Minas Gerais com 8,47% (COMPANHIA NACIONAL DO ABASTECIMENTO, 2012).

O ponto central do agronegócio da cana-de-açúcar baseia-se nos programas de melhoramento genético, sendo que a cultivar o aspecto de maior importância, contribuindo para o crescimento da produtividade média do setor sucroalcooleiro brasileiro e pela melhoria de qualidade da matéria-prima para produção de açúcar e etanol.

O Programa de Melhoramento Genético da Cana-de-Açúcar da Universidade Federal de Viçosa (PMGCA-UFV) visa o desenvolvimento de cultivares através de ações de cooperação entre produtores, entidades de pesquisa e usinas do estado de Minas Gerais. Esse programa está inserido na Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA), que consiste em um convênio firmado entre diversas universidades federais, em que cada uma desenvolve clones em suas regiões com base em sementes produzidas pela Universidade Federal de Alagoas (UFAL) (BARBOSA et al., 2012).

Na fase T1, cruzamentos específicos estabelecidos pela UFAL produzem ao redor de dois milhões de indivíduos, geralmente organizados em famílias sendo realizadas as primeiras seleções de plantas ou de famílias (grupos de indivíduos que possuem ao menos um pai em comum). Por meio de processo de propagação vegetativa, o clone selecionado nessa fase é conduzido para as fases seguintes, onde são plantados em delineamentos com repetições para melhor identificar aqueles potencialmente superiores, visando incluí-los nos experimentos de avaliação (fase FE), em diferentes locais e em anos sucessivos. Após a fase T1 novos materiais não mais são introduzidos, ou seja, os materiais das fases T2 (segunda fase de seleção), T3 (terceira fase de seleção), FE e FM (fase de multiplicação) formam um subconjunto daqueles presentes na fase T1.

A seleção executada em T1 é crucial para o sucesso do programa, pois representa um passo importante na obtenção de novos clones. Quanto mais indivíduos forem avaliados eficientemente nessa fase, maior a probabilidade de sucesso do programa, encontrando-se indivíduos promissores. Tais peculiaridades dessa fase associadas à própria natureza de longa duração do programa de melhoramento da cana-de-açúcar corrobora para a necessidade de métodos que aumentem a eficiência na seleção de novos genótipos.

Apesar da seleção massal ser aplicada rotineiramente nas fases iniciais do programa de melhoramento (BRESSIANI, 2001; MATSUOKA et al., 2005), esse tipo de seleção tem sofrido críticas devido a sua ineficiência (BRESSIANI, 2001; HOGARTH et al., 1997; SKINNER, 1971) em razão da falta de repetição, da competição entre plantas e de efeitos de confundimento produzido pelo ambiente (KIMBENG; COX, 2003). Kimbeng e Cox (2003) ainda afirmam que a seleção de famílias seguida da seleção clonal apresentaria maiores ganhos que a simples seleção isolada, seja somente de clones ou de famílias, principalmente em caracteres de baixa herdabilidade.

Seguindo essa linha de pensamento, alguns programas de melhoramento têm utilizado essa nova estratégia na busca por melhores materiais (BRESSIANI, 2001; COX et al., 1996; KIMBENG; COX, 2003). Tal estratégia se justifica devido a maior probabilidade de se encontrar, nas

progênies, materiais com características favoráveis em famílias de elevados valores genotípicos (BARBOSA et al., 2005; RESENDE; BARBOSA, 2005).

Resende (2002b) mostra que a estratégia de seleção ótima em cana-de-açúcar seria através da predição de valores genotípicos usando o BLUP (*Best Linear Unbiased Predictor*) individual (BLUPI). Este procedimento usa simultaneamente a informação de família e de indivíduos para a seleção. No entanto, esse método dificilmente é usado em programas de melhoramento devido a problemas operacionais relacionados à obtenção dos dados em nível de planta.

Resende e Barbosa (2006) propõem a seleção das famílias com valores genotípicos acima da média geral, seguido da simulação do número de indivíduos a serem selecionados em cada família de acordo com a relação entre os seus valores genotípicos e do número de indivíduo que se deseja selecionar na melhor família, resultando no procedimento denominado BLUP individual simulado (BLUPIS). No entanto, para se usar essa metodologia, é necessário pesar as plantas de todas as parcelas dos experimentos, o que muitas vezes restringe o número de famílias a serem avaliada.

Apesar dos diversos avanços nos procedimentos de seleção de materiais promissores no melhoramento, ainda há a carência de métodos que possam aumentar a eficiência da seleção, fazendo uso de técnicas simples e práticas no processo de seleção de famílias de cana-de-açúcar. As dificuldades da utilização do BLUPIS na seleção de famílias estão relacionadas com o grande volume de informação a ser analisado e com a logística necessária para a coleta e o processamento dos dados em tempo hábil para seleção, tendo em vista que a coleta dos dados é realizada no final do ciclo. Portanto, como forma de contornar o problema de pesagem de todas as parcelas, tem-se buscado alternativas para a coleta de dados em nível de campo, de modo a agilizar o processo de seleção de famílias. Nesse sentido, a definição de classes (categorização) para as variáveis consideradas componentes da produção da cultura (número de colmos, diâmetro de colmos e altura de colmos), se devidamente definidas e validadas experimentalmente, permitiriam grande redução do tempo gasto na coleta de dados.

Uma maneira prática de seleção em cana-de-acúçar seria categorizar os componentes de produção, via de árvores de decisão, especificamente pelo algoritmo CART, que identificaria as famílias com o maior potencial de produção através da combinação de variáveis.

CART (Classification and Regression Trees) são métodos estatísticos não paramétricos utilizados na partição de dados através de determinadas regras específicas desempenhadas por divisões binárias. A metodologia começou a ser desenvolvida em 1973, por Breiman et al. (1984). São essencialmente computacionais e tiveram seu início na grande área das ciências sociais como forma de manipular e interpretar grande volume de dados.

O uso dessa potente ferramenta se justifica não somente como uma forma a mais de interpretar os dados, mas também de possibilitar a leigos o entendimento de seus resultados. O objetivo dessa técnica é explicar a variabilidade da variável dependente em função de variáveis independentes através de divisões binárias (FINCH; SCHNEIDER, 2007).

O algoritmo CART pode construir árvores de classificação e de regressão. As árvores de classificação são construídas quando a variável resposta é categórica enquanto as árvores de regressão são obtidas quando a variável resposta é contínua. Scholes et al. (2011) destacam como vantagem do método o fato do algoritmo avaliar todas os possíveis preditores e divisões. Além disso, o algoritmo pode ser aplicado para outros conjuntos de dados que incluam as mesmas variáveis utilizadas na construção da árvore.

A aplicação do algoritmo é presente em diversas áreas como nas Ciências Sociais (KUCUKKOCAOGLU; ALP, 2012; OZSOY; SAHIN, 2008), nas Biomédicas (GRUBINGER et al., 2010; SCHOLES et al., 2011), na Ecologia (DE'ATH; FABRICIUS, 2000) e nas agrárias (WILLIAMS et al., 2009) esta última tendo número reduzido de trabalhos.

Buscando contribuir para a avaliação de técnicas alternativas de coleta de dados, além de reduzir os custos na fase inicial (T1) dos programas de melhoramento de cana-de-açúcar, este trabalho teve como objetivo sugerir a categorização, por meio do CART, dos componentes de

produção como estratégia de seleção entre e dentro de famílias através da comparação de seu desempenho com os métodos usuais, BLUP e BLUPIS.

2. REFERENCIAL TEÓRICO

A base teórica do CART está apresentada a seguir. Todas essas informações foram retiradas de Breiman et al. (1984).

2.1. Teoria e definições do algoritmo CART

A estrutura de uma árvore de regressão se constitui sempre de divisões binárias. Um nó pai é aquele dá origem a dois nós filhos. Nós que dão origem a valores da variável resposta são chamados nós folha.

Um caso ou uma observação consiste de pares de dados $(\mathbf{x}, y)$. Conceitua-se $\mathbf{x}$ como o vetor medidas ou variável independente e $\mathbf{X}$ como o espaço medidas definido como o espaço contendo todos os possíveis vetores medidas. Seja y o que se conhece usualmente de variável resposta ou dependente.

Uma regra de predição ou preditor é uma função $d(\mathbf{x})$ definida em $\mathbf{X}$ que retira valores reais.

Define-se como amostra de aprendizado aquela constituída de N casos $(\mathbf{x}_1, y_1, \ldots \mathbf{x}_n, y_n)$ utilizada para construir o preditor $d(\mathbf{x})$.

Definimos $R(d)$ como o erro quadrático médio verdadeiro do preditor $d(\mathbf{x})$:

$$R(d) = E\big(Y - d(\mathbf{x})\big)^2 \qquad (1)$$

Em que:

$R(d)$ é o erro esperado usando $d(\mathbf{x})$.

Y é o vetor resposta.

$d(\mathbf{x})$ é o preditor de Y.

2.2. Estimação de erros:

Existem algumas maneiras de se estimar o erro desse preditor. As mais utilizadas são a estimação por V-fold cross validation (validação cruzada), por test sample (amostra teste) e por resubstituição.

A estimação do erro por meio da validação cruzada $\widehat{R^{cv}(d)}$ se baseia na divisão da amostra de aprendizado L em V subconjuntos, cada um contendo o mesmo número de observações, sempre que possível. Para cada v, com $v = 1,2, \ldots, V$ aplica-se o processo de construção para a amostra de aprendizado $L - L_v$. Tomando como preditor $d^{(v)}(\mathbf{x})$ obtém-se:

$$\widehat{R^{cv}(d)} = \frac{1}{N}\sum_v \sum_{(\mathbf{x}_n, y_n)\in L_v}\left(y_n - d^{(v)}(\mathbf{x}_n)\right)^2 \quad (2)$$

Em que:

$\widehat{R^{cv}(d)}$ é o erro estimado por validação cruzada.

$d^{(v)}(\mathbf{x}_n)$ é o preditor em cada $L - L_v$ subconjunto.

$\mathbf{x}_n$ é o vetor medidas.

N é o tamanho da amostra.

y_n é o valor assumido pela variável resposta para cada observação.

A estimação por meio da amostra teste $\widehat{R^{ts}(d)}$ é obtida dividindo a amostra de aprendizado L em L_1 e L_2 (aleatoriamente). A partir de L_1 se obtém o preditor $d(\mathbf{x})$ e L_2 é usado para estimar o erro, tal qual:

$$\widehat{R^{ts}(d)} = \frac{1}{N_2}\sum_v \sum_{(\mathbf{x}_n, y_n)\in L_2}(y_n - d(\mathbf{x}_n))^2 \quad (3)$$

Em que:

$\widehat{R^{ts}(d)}$ é o erro estimado por amostra teste.

N_2 é o tamanho da amostra do subconjunto L_2.

$d(\mathbf{x}_n)$ é o preditor obtido pelo subconjunto L_1.

y_n é o valor assumido pela variável resposta para cada observação.

Em outras palavras, a amostra teste é um caso especial da estimação de erro pela validação cruzada com $V = 2$.

O verdadeiro erro quadrático médio sofre influência da escala da resposta em questão. Pode-se remover essa adversidade trocando o preditor com a média populacional da variável resposta:

$$R(\mu) = E(Y - \mu)^2 \tag{4}$$

Em que:

$R(\mu)$ é o erro quadrático médio usando μ como preditor de Y.

Y é o vetor resposta.

Observa-se que esse erro quadrático médio agora nada mais é do que a própria variância da variável resposta.

Divide-se o erro $R(d)$ por essa variância e obtém-se o erro relativo dado por:

$$RE(d) = \frac{R(d)}{R(\mu)} \tag{5}$$

Essa manipulação matemática permite a comparação do ajuste de árvores de regressão de diferentes escalas. Situação que seria similar à covariância e a correlação de duas variáveis aleatórias. Enquanto a covariância não é comparável, podendo assumindo valores de menos infinito à mais infinito, a correlação assume valores de -1 a +1.

Contudo, deve-se ressaltar que os valores assumidos pelo erro relativo são, normalmente, menores que a unidade. É possível obter erros relativos maiores ou iguais a um, que no caso seriam valores associados à baixa precisão do preditor.

Já que

$$\bar{y} = \frac{1}{N}\sum_n y_n \tag{6}$$

e

$$R(\bar{y}) = \frac{1}{N}\sum_n (y_n - \bar{y})^2 \tag{7}$$

a estimativa do erro relativo por validação cruzada é dada por:

$$RE^{cv}(d) = \frac{R^{cv}(d)}{R^{cv}(\bar{y})} \quad (8)$$

em que:

$R^{cv}(d)$ é o estimador do erro por validação cruzada

$R^{cv}(\bar{y})$ é o estimador do erro utilizando a média amostral da variável resposta como preditor sob 10-fold cross-validation.

Todos esses passos demonstrados previamente levam ao conceito de erro relativo. O erro relativo é a medida padrão de precisão do algoritmo CART para variáveis respostas contínuas (árvores de regressão).

Em uma árvore de regressão, o espaço $\mathbf{X}$ é particionado através de divisões binárias até que se atinja os nós terminais. Cada nó terminal será designado como t e o valor da variável resposta predita como $y(t)$.

Dado uma amostra de aprendizado L, construimos o preditor $d(\mathbf{x})$ e estimamos o erro $R(d)$ através da estimativa por resubstituição. Definimos como estimativa de resubstituição para $R(d)$ a seguinte expressão:

$$R(d) = \frac{1}{N}\sum_n \left(y_n - d(\mathbf{x}_\mathrm{n})\right)^2 \quad (9)$$

Em que:

N é o tamanho da amostra;

y_n são as variáveis respostas;

$d(\mathbf{x}_\mathrm{n})$ é o preditor.

y_n é o valor assumido pela variável resposta para cada observação.

A desvantagem desse método é que o mesmo conjunto de dados que é utilizado para construir o preditor é utilizado para estimar o seu erro. Isso implica em tornar esse método o pior dos três citados até aqui. Algumas demonstrações que forem realizadas para estimativa por resubstituição

podem ser extrapoladas para estimativas por validação cruzada e amostra teste.

O valor que minimiza o erro de estimado por resubstituição é a média de y_n para todas as observações $(\mathbf{x}_1, y_1, \ldots \mathbf{x}_n, y_n)$ em determinado nó terminal t. Dado que:

$$\bar{y} = \frac{1}{N(t)} \sum_{x_n \epsilon t} y_n \tag{11}$$

em que:

$N(t)$ é o tamanho da amostra em determinado nó terminal t.

y_n é o valor assumido pela variável resposta para cada observação.

Obtém-se a estimativa de erro:

$$R(d) = \frac{1}{N} \sum_n (y_n - \bar{y}_n)^2 \tag{10}$$

Mudando a notação de $R(d)$ para $R(t)$, chega-se ao erro estimado de resubstituição para o nó terminal t tomando como valor predito $\bar{y}(t)$.

$$R(t) = \frac{1}{N} \sum_{x_n \epsilon t} (y_n - \bar{y}_n)^2 \tag{12}$$

Em que:

$R(t)$ é o erro estimado de um nó pai.

y_n é o valor assumido pela variável resposta para cada observação.

O erro estimado para cada nó t é a média da soma dos quadrados dos desvios em relação à média. Extrapola-se essa soma então para todos os nós terminais $\tilde{T}$, sendo:

$$R(T) = \frac{1}{N} \sum_{t \epsilon \tilde{T}} \sum_{x_n \epsilon t} (y_n - \bar{y}_n)^2 \tag{13}$$

Que pode ser simplificado como:

$$R(T) = \sum_{t \epsilon \tilde{T}} R(t) \tag{14}$$

Ou seja, a estimativa do erro de toda a árvore é a soma das estimativas dos erros de cada nó terminal t.

Sabendo que um conjunto de possíveis divisões s em um determinada divisão é dada por S. A melhor divisão s^* será aquela que reduzir ao máximo o erro de predição. Isto é:

$$\Delta R(s^*, t) = max_{s \epsilon S} \Delta R(s, t) \quad (15)$$

Em que:

$\Delta R(s, t) = R(t) - R(t_1) - R(t_2)$;

s é uma possível divisão qualquer.

$R(t_1)$ e $R(t_2)$ são os erros estimados dos nós filhos.

Apesar de toda teoria de CART, o pacote `rpart()` possui algumas distinções do método. A principal delas é que o erro em cada nó é simplesmente a soma de quadrados ao invés da soma de quadrado médio conforme definido em (13). Assim, o erro relativo passa a ser a razão da soma de quadrados entre determinada árvore e a árvore raiz (árvore sem nenhum nó). Temos assim:

$$R(t) = \sum_{x_n \epsilon t} (y_n - \bar{y}_n)^2 \quad (17)$$

$$R(T) = \sum_{t \epsilon \tilde{T}} R(t) \quad (18)$$

$$\Delta R(s^*, t) = max_{s \epsilon S} \Delta R(s, t) \quad (19)$$

Em que:

$\Delta R(s, t) = R(t) - R(t_1) - R(t_2)$;

s é uma possível divisão qualquer.

$R(t)$ é o erro estimado de um nó pai.

$R(t_1)$ e $R(t_2)$ são os erros estimados dos nós filhos.

y_n é o valor assumido pela variável resposta para cada observação.

O que o pacote `rpart()` do software R faz é procurar, iterativamente, a divisão que minimizará o erro de cada nó. Uma divisão que consegue reduzir o erro nada mais é do que uma divisão que consegue dividir o vetor de medidas gerando altos e baixos valores de média da variável resposta.

2.3. Seleção de modelos

A poda das árvores de regressão é uma técnica que tem como objetivos reduzir o efeito de overfitting, gerar árvores que sejam eficazes nas classificações de outros conjuntos de dados e facilitar a interpretação de resultados. Breiman et al. (1984) apresentam o 1-SE rule que seleciona o menor valor de 10-fold cross-validation de todas as árvores possíveis, adiciona-se o seu respectivo erro padrão e encontra-se aquela que possui o 10-fold cross-validation imediatamente inferior a esse valor. Faraway (2006) sugere a poda com base simplesmente nos valores de 10-fold cross-validation. Dessa forma, após ter sido criada uma árvore suficiente grande, escolhe-se aquela sub-árvore que possui o menor valor de validação cruzada.

3. MATERIAL E MÉTODOS

3.1. Material vegetal

Foram utilizadas sementes de 110 famílias de irmãos completos, provenientes de cruzamentos realizados na Estação Experimental da Serra do Ouro, da Universidade Federal de Alagoas, localizada no Município de Murici, Estado de Alagoas, no ano de 2006 (BARBOSA et al., 2002). Os cruzamentos foram controlados, sendo denominados biparentais ou cruzamentos simples.

As sementes originadas destes cruzamentos foram coletadas e acondicionadas em embalagens adequadas, e em seguida, enviadas ao

Centro de Pesquisa e Melhoramento de Cana-de-açúcar (CECA), da Universidade Federal de Viçosa, localizado no município de Oratórios, Minas Gerais, com latitude 20°25'S; longitude 42°48'W; altitude 494 m; solo LVE. As mesmas foram semeadas em bandejas com substrato para produção de mudas e acondicionadas em casa-de-vegetação para germinação e crescimento das plântulas. Posteriormente, as plântulas foram transplantadas e ficaram 120 dias sob condição de sombrite com 50% de luminosidade. Finalmente ficaram mais 30 dias sem sombrite para aclimatação.

As plântulas originadas dos cruzamentos biparentais, após aclimatação, foram enviadas para a instalação do experimento de famílias de irmãos completos em uma área experimental do CECA.

No total, 5 experimentos foram instalados em maio de 2007 no delineamento em blocos casualizados, sendo cada experimento constituído de 5 blocos, 22 famílias e 2 testemunhas em comum (variedades comerciais). Cada parcela foi constituída por 20 plantas, distribuídas em dois sulcos de 5 m de comprimento, espaçados em 1,40 m, totalizando 12.000 plantas. As mudas das testemunhas foram oriundas de uma gema plantada em viveiro de forma semelhante às mudas oriundas de sementes.

3.2. Coleta de dados

Em 2009, foram avaliados os caracteres: altura de colmos (AC) em metros, mensurando-se um colmo de cada touceira, desde a base até a primeira folha cuja seção compreendida entre o limbo foliar e a bainha esteja visível; diâmetro de colmos (DC) em centímetros, mensurando-se com paquímetro digital o terceiro internódio, contado da base do colmo para o ápice, de um colmo por touceira e número total de colmos por parcela (NC).

A massa total da parcela (MTP) em kg foi determinada a partir da pesagem de todos os colmos com auxílio de um dinamômetro. A produtividade de colmos em tonelada de cana por hectare (TCH) foi determinada pela fórmula:

$$TCH = \frac{MTP \times 10000}{AP \times 1000} \qquad (20)$$

em que MTP é a massa total da parcela em Kg e AP é a área da parcela em m^2 (cada parcela tinha 14 m^2).

3.3. Análise dos dados

3.3.1. Seleção via CART

Especificamente nesse trabalho, trabalhou-se com as árvores de regressão já que o intuito foi de criar classes das três variáveis, AC, DC e NC, para os TCH's verificados nos cinco experimentos conjuntos, equivalendo a um delineamento em blocos aumentados.

Para a construção das árvores de regressão, utilizou-se os dados coletados referentes somente às testemunhas para as variáveis altura média de colmos, diâmetro médio de colmos, números totais de colmos por parcela e tonelada de cana por hectare, totalizando 1000 observações.

As árvores de regressão não podem ser geradas caso o número de observações seja muito reduzido. Sendo assim, a simulação anterior à utilização do algoritmo teve como objetivo prover de solução uma situação como essa.

Através da decomposição de Cholesky, a partir das matrizes de covariância $\sum$ (positiva definida) do experimento, foi possível obter $\sum = \mathrm{CC}'$, em que C é uma matriz triangular inferior mxm, denominada fator de Cholesky. Foi simulado então um vetor multivariado normal $\mathrm{X} = \mu + \mathrm{CZ}$, em que μ é o vetor de médias das testemunhas do experimento, C é o fator de Cholesky proveniente da matriz de variância e covariâncias e Z vetor de variáveis aleatórias IID com $\mathrm{N}(0,1)$. Com base nesse procedimento e através de 1000 iterações, os dados gerados foram submetidos ao procedimento padrão do algoritmo CART.

Para se obter estimativas mais precisas, recorreu-se à utilização de podas das árvores. Através das observações das testemunhas, procedeu-se à obtenção das árvores de regressão com dados simulados, árvores de regressão com dados não simulados, podadas (segundo o 10-fold cross-validation e o 1-SE rule) e não podadas. Mesmo após as podas, mecanismo

que permite extrapolar as categorizações para outros conjuntos de dados, que no caso desse trabalho seriam outros experimentos, essa não é prática prudente já que o efeito de ambiente em cada um pode variar. Assim sendo, o procedimento de poda teve como único objetivo a obtenção de estimativas mais precisas.

Com a intenção de se obter um ponto de corte de seleção de clones, localizaram-se, nas árvores obtidas, combinações de variáveis que originassem produções (TCH) superiores às médias das duas testemunhas. Através dessas combinações, foram selecionadas as famílias que apresentassem os valores desejados para NC, DC e AC.

As famílias selecionadas foram separadas em 3 classes visando definir o número de indivíduos a serem selecionados em cada uma das famílias indicadas pelo algoritmo CART. As classes foram definidas com base no número de repetições em que a família foi selecionada pelo CART. Dessa maneira, a primeira classe foi formada por famílias selecionadas pelo CART em cinco repetições, a segunda classe foi composta pelas famílias selecionadas em quatro repetições e a terceira classe composta por famílias selecionadas em três repetições. Foram selecionados na melhor classe 30% de indivíduos em cada família, seguido de 20 e 10% dos indivíduos de cada família.

3.3.2. Seleção via BLUP e BLUPIS

Os dados de TCH foram analisados via modelos mistos REML/BLUP, usando um modelo estatístico associado à avaliação de genótipos, no delineamento em blocos incompletos com média da parcela, considerando a equação matricial descrita abaixo, conforme Resende (2002):

$y=Xr+Zg+Wb+e$, em que: y é o vetor de dados; r é o vetor de efeitos de repetição somados a média geral (assumidos como fixos); g é o vetor de efeitos genotípicos (assumidos como aleatórios); b é o efeito de blocos incompletos (aleatórios); e é o vetor de erros (aleatórios) e X, Z e W representam as matrizes de incidência, respectivamente, para os efeitos de r, g e b.

No procedimento BLUP a seleção foi realizada seguindo a estratégia usada pelo programa de melhoramento da Austrália (STRINGER et al., 2011), com a seleção de 40% das famílias avaliadas. As famílias a serem selecionadas foram separadas em 4 classes com base nas médias de TCH. Cada classe foi composta por 11 famílias, sendo selecionados, 40% dos indivíduos em cada família da primeira classe e 30, 20 e 10% dos indivíduos em cada família, respectivamente, nas classes 2, 3 e 4.

Pelo procedimento BLUPIS foram selecionadas as famílias que apresentaram médias de TCH superiores a média geral. O número de indivíduos selecionados em cada família k foi calculado por $n = (\hat{g}_k/\hat{g}_j)n_j$, em que $\hat{g}_j$ refere-se ao valor genotípico da melhor família e n_j equivale ao número de indivíduos selecionados na melhor família. Resende e Barbosa (2006), nesse trabalho n_j foi igual a 27. As análises via modelos mistos foram obtidas através do software SELEGEN-REML/BLUP (RESENDE, 2007).

3.3.3. Comparação entre BLUP, BLUPIS e CART

Gerou-se uma matriz de confusão para cada árvore, com os valores de acurácia, false positive rate e precisão de modo a facilitar a visualização das correspondências e distinções de seleção entre o BLUP, BLUPIS e o CART.

A medida de Acurácia se refere ao número de famílias, selecionadas e não selecionadas, quando comparados com BLUP ou BLUPIS dividido pelo total de famílias do experimento. False Positive Rate se refere ao número de famílias que o CART selecionou, erroneamente, quando comparado com BLUPIS ou BLUP, dividido pelo total de selecionados por estes últimos. E, finalmente, Precisão é o número de famílias selecionadas, simultaneamente, pelo CART, BLUPIS ou BLUP dividido pelo total de famílias selecionado pelo BLUP ou BLUPIS.

Todas as análises e gráficos do algoritmo CART foram obtidos através do software livre R (R DEVELOPMENT CORE TEAM, 2012), especificamente no pacote `rpart()`.

4. RESULTADOS E DISCUSSÃO

Depois de geradas as árvores de regressão, foram definidas as classes relativas às variáveis NC, DC e AC mais associadas à TCH`s superiores às testemunhas.

Através do BLUPIS para a variável TCH, foram selecionadas cinquenta e duas famílias. A Tabela 1 dispõe o número de indivíduos selecionados pelo BLUP, BLUPIS, CART com dados sem e com simulação. Enquanto as famílias foram ranqueadas com base nas médias genotípicas de TCH pelos procedimentos BLUP e BLUPIS, as famílias selecionadas pelo CART foram ranqueadas com base no número de blocos em que cada uma foi indicada para seleção.

Com a intensidade de seleção aplicada nesse trabalho, enquanto a seleção via BLUP indicou 40 indivíduos a serem selecionados na melhor família, no BLUPIS foram indicados 27, enquanto o CART indicou 30. No total, foram indicados para seleção 1100 pelo BLUP, 1077 pelo BLUPIS, 1022 pelo CART com dados sem simulação e 890 pelo CART a partir de dados simulados (Tabela 1).

Tabela 1 – Médias genotípicas (u+g) de TCH das famílias selecionadas via BLUP, BLUPIS e CART, usando dados com simulação e sem simulação, número de repetições em que cada família foi selecionada pelo CART (Rep) e número de indivíduos selecionados dentro de cada família (nk).

Ordem	Sem simulação						Com simulação		
	Família	u+g	Rep	n_k			Família	Repetição	n_k
				BLUP	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

(Continua)

(Continuação)

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
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

(Continua)

(Continuação)

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	**Total**		890
Total				1100	1077	1020			

* Em negrito as famílias não selecionadas pelo CART

A Tabela 2 exibe as matrizes de confusão entre CART, BLUP e BLUPIS e as respectivas medidas de Acurácia, False Positive Rate e Precisão. O uso de tantas medidas se justifica, pois na interpretação das matrizes de confusão, um parâmetro apenas não é capaz de retratar todo o desempenho do classificador em questão. Para o cenário específico de seleção de famílias em de cana-de-açúcar, melhor será o desempenho do CART quanto maior for a Acurácia e menor for o False Positive Rate. Em outras palavras, quanto maior for o número de predições corretas de famílias selecionadas e não selecionadas, e, menor for o número de famílias selecionadas erroneamente, melhor será o desempenho do algoritmo.

Para dados não simulados, a Tabela 2 mostra que das 52 famílias selecionadas pelo BLUPIS, o CART foi capaz de identificar 38. No entanto, o CART selecionou 14 famílias a mais. A Acurácia do CART em relação ao BLUPIS ou BLUP foram equivalentes, apresentando valor de 0,745. Isso, em termos práticos, nos informa que o CART conseguiu prever 74,5% das famílias que foram selecionadas e das que não foram selecionadas. Quando se analisa os valores de False Positive Rate do CART em relação ao BLUPIS ou BLUP, verificaram-se os valores de 0,269 e 0,409. Esses valores representam as porcentagens (26,9% e 40,9%) do total de famílias selecionadas, respectivamente, pelo BLUPIS e BLUP, que o CART selecionou, mas não deveria ter selecionado. E, por último, a precisão do CART, em relação ao BLUPIS ou BLUP, foi, de 0,692 e 0,727, ou seja, do total de famílias selecionadas pelo BLUPIS e BLUP, o CART, selecionou, respectivamente, 69,2% e 72,2% destas.

Tabela 2 – Matrizes de confusão entre as estratégias de seleção de famílias CART, BLUPIS e BLUP, acompanhado das medidas de Acurácia, True Positive Rate, True Negative Rate e Precisão para os dados sem simulação e com simulação.

Sem simulação						
CART	BLUPIS		Total	BLUP		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
Acurácia	0,745			0,745		
False Positive Rate	0,269			0,409		
Precisão	0,731			0,773		
Com simulação						
CART	BLUPIS		Total	BLUP		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
Acurácia	0,736			0,736		
False Positive Rate	0,250			0,386		
Precisão	0,692			0,727		

* S = famílias selecionadas, N = famílias não selecionadas.

Quando se considerou os dados simulados, o CART selecionou 36 das 52 famílias indicadas pelo BLUPIS, e outras 13 famílias. As Acurácias do CART em relação ao BLUPIS ou BLUP foram equivalentes, com o valor de 0,736. Novamente, em termos práticos, esses dados nos informam que o CART conseguiu prever 73,6% das famílias que foram selecionadas e das que não foram selecionadas. Analisando os valores de False Positive Rate do CART em relação ao BLUPIS ou BLUP, verificaram-se os valores de 0,250 e 0,386. Ou seja, 25,0% e 38,6% das famílias que o BLUPIS e o BLUPS selecionaram, respectivamente, foram selecionadas pelo CART quando na verdade não deveriam ser selecionadas. O False Positive Rate, nada mais é do que o erro de seleção de famílias do CART quando comparado com os métodos usuais, BLUPIS e BLUP. A medida de precisão do CART, em relação ao BLUPIS ou BLUP, foi respectivamente de 0,692 e 0,727. Estes nos informam que, do total de famílias selecionadas pelo

BLUPIS e BLUP, o CART, selecionou, respectivamente, 69,2% e 72,2% das mesmas.

O uso da simulação de dados anteriormente ao procedimento CART tem como principal vantagem a possibilidade de simular médias para famílias tidas como ideais, a critério do pesquisador, de forma a definir quais famílias selecionar dentre as presentes em determinado experimento com ausência de clones e/ou variedades testemunhas. A Tabela 2 sugere que o procedimento de simulação foi pertinente visto que as medidas de Acurácia, False Positive Rate e Precisão para dados simulados e não simulados foram bem próximas, indicando uma manutenção do desempenho do algoritmo.

Apesar de ter se recorrido às podas das árvores de forma a se obter estimativas mais precisas, não houve alteração das árvores seja pelo 10-fold cross-validation ou pelo 1-SE rule, tanto para os dados simulados quanto para os dados não simulados. A não alteração das árvores com o procedimento de podas ocorreu porque o algoritmo foi capaz de chegar à árvore ótima, por si só, sem qualquer ajuste do modelo.

A Figura 1 mostra a árvore de regressão com dados não simulados gerada pelo CART. A média de produção das testemunhas foi de 145,81 toneladas de colmo por hectare e produções acima desse TCH foram originadas de famílias com NC acima de 110,5. Isso equivale a dizer que o NC foi categorizado em duas classes, a primeira composta de famílias que apresentam valores de número total de colmos por parcela inferior a 110,5 e a segunda composta de famílias que apresentam esses valores superiores a 110,5 (Tabela 3).

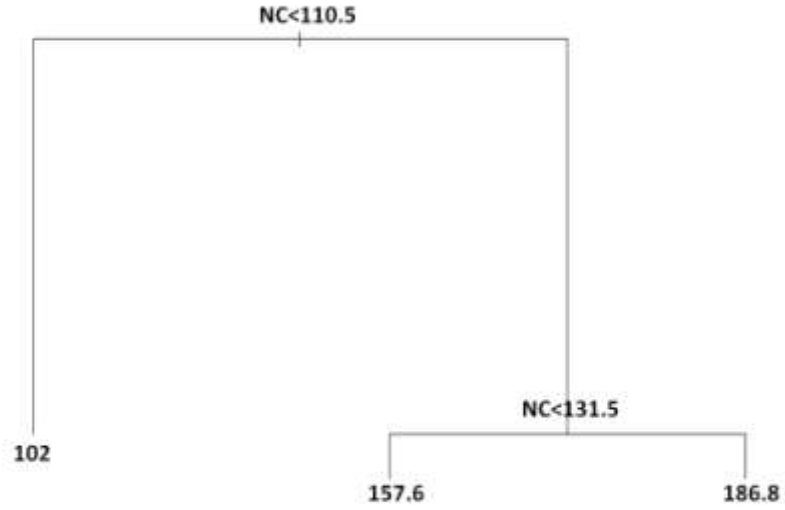

Figura 1 – Árvore de regressão gerada pelo algoritmo CART para os dados das testemunhas, em que NC representa o número de colmos total por parcela não simulado e os nós terminais produções preditas em toneladas de colmo por hectare (TCH).

Tabela 3 – Classes definidas pelas árvores de regressão e as respectivas produções preditas (TCH) para dados simulados e não simulados.

Dados não simulados		
Classe	NC	TCH
1	<110,5	102
2	>110,5	>157,6
Dados simulados		
1	<113,4	<135,6
2	>113,4	>157,0

A Figura 2 representa a árvore com dados simulados. Nessa árvore, produtividades acima de 145,81 toneladas por hectare foram geradas por famílias com número total de colmos por parcela acima de 113,4. Isso, da mesma forma que anteriormente, equivale a dizer que o NC foi categorizado em duas classes sendo primeira composta de famílias que apresentam valores de número total de colmos por parcela inferior a 113,4 e a segunda composta de famílias que apresentam esses valores superiores a 113,4 (Tabela 3).

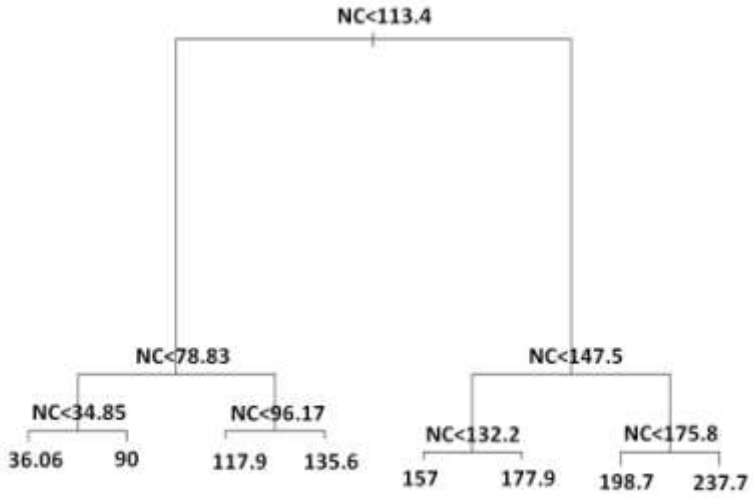

Figura 2 – Árvore de regressão gerada pelo algoritmo CART para os dados das testemunhas, em que NC representa o número de colmos total por parcela simulado e os nós terminais produções preditas em toneladas de colmo por hectare (TCH).

Apesar de todos os componentes de produção terem sido utilizados para gerar as árvores de regressão, o CART descartou os componentes altura de colmos e diâmetro de colmos na predição dos valores de TCH. Isso indica que número de colmos é a única variável eficaz em reduzir as somas de quadrados dos desvios de cada nó e, nesse experimento, seria a única variável determinante do TCH.

O número de colmos tem apresentado o maior efeito direto sobre TCH em análises de trilha com o uso de correlações genotípicas, genéticas e fenotípicas conforme observado por Espósito et al. (2012), Silva et al. (2009) e Sukhchain, Sandhu e Saini (1997) na avaliação de famílias de cana-de-açúcar. Esses autores tem indicado a possibilidade de sucesso na seleção de clones para TCH com base em NC apenas, pois esse é o principal determinante das variações em TCH.

Embora a média da população selecionada pelo CART tenha sido inferior ao BLUP e BLUPIS, tanto nos dados simulados quanto nos não simulados (Tabela 3), a grande vantagem do CART seria a ausência da pesagem de toda a parcela, pois somente com a contagem do número de

colmos seria possível direcionar a seleção das melhores famílias. Obviamente, o procedimento correto de seleção deveria ser o descrito até aqui, não podendo se afirmar que, para outros experimentos, o algoritmo consideraria a variável número de colmos como a única determinante da produção.

Tabela 4 – Média da população selecionada (M_s) em toneladas de colmos por hectare (TCH) e número de famílias selecionadas (n_f) pelas estratégias de seleção via BLUP, BLUPIS e CART.

Estratégia	BLUP	BLUPIS	CART	
			SS	CS
M_s	123,4621	120,4744	115,5319	115,8233
n_f	44	52	52	49

SS = sem simulação; CS = com simulação.

A Tabela 4 mostra a matriz de correlação, do experimento, entre as quatro variáveis, AC, DC, NC e TCH, demonstrando uma alta correlação entre NC e TCH, deixando claro o porquê de, nesse experimento, o NC ter sido considerado pelo algoritmo o mais influente no TCH.

Tabela 5 – Matriz de correlação entre as variáveis; Altura média de colmos por parcela (AC), diâmetro médio de colmos por parcela (DC), número de colmos por parcela (NC) e toneladas de colmos por hectare (TCH).

Variável	**AC**	**DC**	**NC**	**TCH**
AC	1,000*	0,280	0,261	0,355
DC		1,000*	-0,275	-0,071
NC			1,000*	0,931*
TCH				1,000*

* significativo a 5% de probabilidade, pelo teste t.

Além disso, o CART pode ser aplicado na seleção entre e dentro de famílias em experimentos sem delineamento experimental, tendo em vista que a maior parte das famílias avaliadas na primeira fase de teste (T1) são levadas a campo em experimentos sem repetição, com cada sulco representando uma determinada família referência.

Como tem sido recomendada a seleção de famílias seguida da seleção individual de clones, a seleção das melhores famílias em cana-de-açúcar pode ser realizada com base no número de colmos, cuja herdabilidade baseada nas médias de famílias tem sido superior a herdabilidade com plantas individuais, conforme demonstrado por Barbosa et al. (2005) e Kimbeng e Cox (2003). Além disso, o seu efeito direto sobre tonelada de cana por hectare (TCH) conforme verificado em trabalhos de análise de trilha e nos resultados das podas geradas pelo algoritmo CART é de alta magnitude. Essa prática pode levar a uma diminuição de tempo e custos no processo de avaliação e seleção entre e dentro de famílias, aumentando a eficiência desse processo nas fases iniciais dos programas de melhoramento da cana-de-açúcar, uma vez que a seleção indireta através de caracteres menos complexos com maior herdabilidade e de fácil mensuração, pode resultar em maiores progressos genéticos em relação ao uso da seleção direta.

Portanto, a estratégia de seleção via CART pode ser aplicada visando à diminuição dos custos operacionais, pois requer menor quantidade de mão-de-obra e menor tempo de execução, tanto na instalação dos experimentos como na avaliação das famílias, aumentando a eficiência do processo de seleção individual nas fases iniciais dos programas de melhoramento genético da cana-de-açúcar.

5. CONCLUSÃO

O algoritmo CART foi eficiente em definir as classes dos componentes de produção seguido da seleção das melhores famílias no campo com acurácia média próxima de 73% quando comparado com o BLUPIS e BLUP.

6. REFERÊNCIAS

BARBOSA, M. H. P et al. Estratégias de melhoramento genético da cana-de-açúcar em universidades. In: IV Simpósio de atualização em genética e melhoramento de plantas, 2005, Lavras. **Seleção recorrente no melhoramento de plantas**, Lavras : UFLA, 2005. v. 1. p. 43-58.

BARBOSA, M.H.P.; SILVEIRA, L.C.I. Metodologias de seleção, progressos e mudanças no programa de melhoramento genético da cana-de-açúcar da Universidade Federal de Viçosa. **STAB-Açúcar, Álcool e Subprodutos**, v.18, p. 30-32. 2000.

BARBOSA, M. H. P.; SILVEIRA, L.C.I.. Melhoramento Genético e Recomendação de Cultivares. In: SANTOS, F; BORÉM, A,; CALDAS, C. (Org.). **Cana-de-Açúcar:** Bioenergia, Açúcar e Álcool**.** Tecnologias e Perspectivas. Viçosa: Suprema Gráfica e Editora Ltda, 2009, v. 1, p. 313-331.

BARBOSA, M.H.P et al. Genetic improvement of sugar cane for bioenergy: the Brazilian experience in network research with RIDESA. **Crop Breeding and applied biotechnology**, S2, p. 87-98. 2012.

BARBOSA, G.V.S. et al. A brief report on sugarcane breeding program in Alagoas, Brazil. **Crop Breeding and Applied Biotechnology**, v. 2, n. 4, p. 613-616. 2002.

BREIMAN, L. et al. **Classification and Regression Trees.** Boca Raton: Chapman & Hall/CRC, 1984. 358 p.

BRESSIANI, J.A. **Seleção seqüencial em cana-de-açúcar.** 2001. 67 f. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola superior de Agricultura Luiz de Queiroz, Piracicaba, SP, 2001.

CESNIK, R.; MIOCQUE J. **Melhoramento da cana-de-açúcar**. Brasília: Embrapa Informação Tecnológica, 2004. 307 p.

COMPANHIA NACIONAL DO ABASTECIMENTO. **Acompanhamento da safra brasileira:** Cana-de-açúcar. Terceiro levantamento. Brasília, DF, 2012. 18 p.

CONFEDERAÇÃO NACIONAL DAS INDÚSTRIAS. **Bioetanol:** O futuro renovável. Encontro da indústria para sustentabilidade. Brasília, DF, 2012. 81 p.

COX, M.C. et al. Family selection improves the efficiency and effectiveness of a sugarcane improvement program. In: Wilson, J.R. et al. (eds.). **Sugarcane:** Research Towards Efficient and Sustainable Production. Brisbane: CSIRO Division of Tropical Crops and Pasture, 1996. p. 42-43.

DE'ATH, G; FABRICIUS, K.E. Classification and regression trees: A powerful yet simple techinique for ecological data analysis. **Ecology,** v. 81, n.11, p. 3178-3192, nov. 2000. Disponível em: <http://www.jstor.org/stable/177409>. Acesso em: 25 jul. 2012.

ESPÓSITO, D.P. et al. Análise de trilha usando valores fenotípicos e genotípicos para componentes do rendimento na seleção de famílias de cana-de-açúcar. **Ciência Rural**, v. 42, n. 1, p. 38-44, jan. 2012.

FARAWAY, J.J. **Extending the linear model with R.** Generalized linear, Mixed effects and nonparametric regression. Boca Raton: Chapman & Hall/CRC. 2006. 317 p.

FINCH, H; SCHNEIDER, M.K. Classification Accuracy of Neural Networks vs. Discriminant Analysis. **Methodology**, v. 3, n. 2, p. 47-57. 2007.

GRUBINGER, T; KOBEL, C; PFEIFFER, K. Regression tree construction by bootstrap: Model search for DRG-systems applied to Austrian health-data.

BMC Medical Informatics and decision making, v. 10, n. 9. 2010. Disponível em: <http://www.biomedcentral.com/1472-6947/10/9>. Acesso em: 17 jul. 2012.

HOGARTH, D. M.; COX, M. C.; BULL, J. K. Sugarcane improvement: Past achievements and future prospects. In: Kang, M.S. **Crop Improvement for the 21st century**. Baton Rouge: Louisiana State University, 1997. p. 29-56.

KIMBENG C.A.; COX, M.C. Early generation selection of sugarcane families and clones in Australia: a review. **Journal American Society of Sugarcane Technologists**, v. 23, p. 20-39. 2003.

KUCUKKOCAOGLU, G; ALP, O.S. IPO mechanism selection by using Classification and Regression Trees. **Qual Quant.** v. 46, p. 873–888. 2011.

MATSUOKA, S.; GARCIA, A.A.F.; ARIZONO, H. Melhoramento da cana-de-açúcar. In: Borém, A. 2. ed. **Melhoramento de espécies cultivadas**. Viçosa: UFV, 2005. p. 225-274.

NEVES, M.F; CONEJERO, M.A. Sistema agroindustrial da cana: cenário e agenda estratégica. **Economia aplicada**, São Paulo, SP, v. 11, n. 4, p. 587-604, out./dez. 2007.

OZSOY, O; SAHIN, H. Housing price determinants in Istanbul, Turkey: An application of the classification and regression tree model. **International Journal of Housing Markets and Analysis**, v.2, n. 2, p. 167-178. 2009. Disponível em: <http://dx.doi.org/10.1108/17538270910963090>. Acesso em: 25 Jul, 2012.

R DEVELOPMENT CORE TEAM. **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Áustria, 2010. Disponível em:<http://www.R-project.org>. Acesso em: 10 jan. 2012.

RESENDE, M.D.V. **Genética biométrica e estatística no melhoramento de plantas perenes**. Brasília: Embrapa Informação Tecnológica, 2002b. 975 p.

RESENDE M.D.V. **Selegen-REML/BLUP:** sistema estatístico e seleção genética computadorizada via modelos lineares mistos. Colombo: Embrapa Florestas, 2007. 359 p.

RESENDE, M. D. V; BARBOSA, M. H. P. **Melhoramento genético de plantas de propagação assexuada**. Colombo: Embrapa Informação Tecnológica, 2005. 130 p.

RESENDE, M. D. V.; BARBOSA, M. H. P. Selection via simulated Blup based on family genotypic effects in sugarcane. **Pesquisa Agropecuária Brasileira**, Brasília, DF, v. 41, n. 3, p. 421-429. 2006.

SCHOLES, D. et al. Improving automated case finding for ectopic pregnancy using a classification algorithm. **Human Reproduction**, Oxford, v. 26, n. 11, p. 3163-3168. 2011.

SILVA, F.L. et al. Análise de trilha para os componentes de produção de cana-de-açúcar via blup. **Ceres**, v. 56, n. 3, p. 308-314, mai./jun. 2009.

SKINNER, J. C. Selection in sugarcane: a review. Proceedings **International Society Sugarcane Technologists**, v. 14, p. 149-162. 1971.

STRINGER, J.K et al. Family Selection Improves the Efficiency and Effectiveness of Selecting Original Seedlings and Parents. **Sugar Tech**, v. 13, n. 1, p. 36-41. 2011.

SUCKHCHAIN; SANDHU, D.; SAINI, G.S. Inter-relationships among cane yield and commercial cane sugar and their component traits in autumn plant crop of sugarcane. **Euphytica**, v. 95, p. 109-114. 1997.

WILLIAMS, M.M et al. Linkages among agronomic, environmental and weed management characteristics in North American sweet corn. **Field Crops Research**, v. 113, p. 161–169. 2009.