SUGARCANE FAMILIES SELECTION IN EARLY STAGES BASED ON CLASSIFICATION BY DISCRIMINANT LINEAR ANALYSIS

Édimo Fernando Alves MOREIRA[1]
Luiz Alexandre PETERNELLI[1]

- *ABSTRACT: A major challenge in breeding programs is the efficient selection of genotypes in the early stages. The efficiency of selection in these phases is critical for the program targets, since, due to the particularity of sugarcane, the genotypes selected in the early stages will be assessed in later stages. The objective of this study was to compare selection by linear discriminant analysis with family selection based on estimates of the variable tons of cane per hectare (TCHe), defined by the indirect traits number of stalks, stalk diameter and stalk height, as alternatives to the selection of promising sugarcane families. Also simulations were considered in order to augment the training observations before analysis. Five different simulation scenarios were considered: without simulation and with 500, 750, 1000, or 2000 families simulated. The methods were compared and evaluated by the apparent error rate (AER). Linear discriminant analysis indicated a high concordance with the selection based on the measured, or real, TCH (TCHr) and can be used for early selection of sugarcane families. In the simulated scenarios, results from selection based on linear discriminant analysis were better than those of selection for TCHe. The AER is minimized when 1,000 families are simulated to augment the training observations.*

- *KEY WORDS: Simulation; plant breeding; Saccharum spp.*

1 Introduction

Introduced in the colonial period, sugarcane (*Saccharum* spp.) went on to become one of the most important crops for the Brazilian economy. Brazil is the world's largest producer of sugar and ethanol and is constantly expanding to more foreign markets, due to the use of biofuels as alternative energy (BARBOSA and SILVEIRA, 2010).

According to Waclawovsky et al. (2010), the world average yield of sugarcane is about80 t/ha and the theoretical maximum yield 380 t/ha. In the growing season 2014/2015, the nation-wide mean yield was 72.444 t/ha (CONAB, 2014). Therefore, Brazilian cultivars produce far less than the genetic potential of the crop, indicating the need for breeding.

A major challenge in sugarcane breeding programs is the selection of genotypes in early stages. Efficient selection in these stages is extremely relevant for the goals of these programs, since the selected genotypes are tested in later stages.

[1] Universidade Federal de Viçosa - UFV, Centro de Ciências Exatas, Departamento de Estatística, Av. P H Rolfs, S/N, CEP: 36570-000, Viçosa, Minas Gerais, Brasil. E-mail: *edimo.fernando@gmail.com; peternelli@ufv.br*

Most commonly, mass selection is used for early selection in sugarcane breeding programs. This method can be summarized as the individual visual selection of clones for traits correlated with sugarcane yield and health (BARBOSA and SILVEIRA, 2010).

Stringer et al. (2011) proposed the selection of families instead of that of individual clones, followed by selection of the best genotypes within the best families, so that the heritability of yield- related traits in families is higher than in individual plants. Thus, it is preferable to prioritize the selection of promising families followed by individual selection of clones in the best families.

The statistical methods used in the early stages of sugarcane selection are the BLUP individual (Best Linear Unbiased Predictor) (BLUPI) (RESENDE, 2002) and individual simulated BLUP (BLUPIS) methods (RESENDE and BARBOSA, 2006). In the case of balanced data, as in this study, selection for the main variable can be simplified. In this case, only families with better performance than the overall mean for the actual or measured, tons of cane per hectare (TCHr) have to be taken into consideration.

The drawback of these methods is the need to weigh all main plots to determine TCHr, representing a major operational problem.

A common alternative to circumvent the problem of weighing in the field is to estimate the variable tons of cane per hectare (TCH), to select those with TCH above the overall mean. This estimate is obtained by the indirect traits number of stalks (NS), stalk diameter (SD) and stalk height (SH), commonly used to evaluate sugarcane yield (CHANG and MILLIGAN, 1992).

Another technique to avoid the problem of weighing sugarcane is discriminant analysis. This has been widely used in breeding programs for classification (NOGUEIRA et al., 2008; SUDRÉ et al., 2006). The method consists of functions established to classify an observation x, based on measurements of a number p of traits of this observation, in one of several populations $\prod_i$, with different $i = 1,2,...,n$ (KHATTREE and NAIK, 2000).

In this context, the variables NS, SD and SH of some selected and unselected families can be provided, based on TCHr selection, and a linear discriminant function can be established to include new observations in one of these groups. This would reduce the field work, optimizing the process of selecting the best families.

The objective of this study was to use discriminant analysis to classify families in selected or unselected families, based on the indirect traits NS, SD and SH, and to compare this with TCHe selection. Additionally, the use of simulation to augment the training population will be evaluated as a way to reduce the apparent error rate.

2 Material and methods

2.1 Plant material and data set

The data used were the results of five experiments, carried out at the Center for Research and Improvement of sugarcane (ECSC), of the Federal University of Viçosa, located in the municipality of Oratory, Minas Gerais (20°25'S, 42°48'W, 494 m in altitude). The experiments were set up in a randomized block design with 5 replications and 22 families each. The experimental units consisted of 20 plants in two 5-m long rows, spaced 1.40 m apart.

The following traits were evaluated: stalk height (SH) in meters, of one stalk per stool, from the base to the first visible dewlap; stalk diameter (SD) in centimeters, measured with a digital caliper at the third internode, counted from the base of the stalk to the apex; total number of stalks per plot (NS) and real tons of cane per hectare (TCHr), measured by weighing the stalks of each plot.

The best families, with a TCHr above the overall experimental mean, were selected by weighing all families of each of the five experiments.

Thus, the data set contained values of SD, SH and NS and the result of selection for TCHr in five different scenarios: no simulation, and simulation of values for 500, 750, 100 and 2000 families.

2.2 Simulation

Since 110 families could be insufficient to estimate discriminant functions with wide generalization ability, values of NS, SH, SD and TCHr, with a multivariate normal distribution for 500, 750, 1000 and 2000 families were simulated. Simulations were based on the structure of means and covariance obtained from the 22 families of experiment 1.

To perform the simulation data, a vector of means and a matrix of covariances were obtained for each variable from experiment 1. Simulation was performed using the Cholesky decomposition in the correlation matrix of the variables NS, SD, SH, and TCHr. This method is widely used for simulation with multiple correlated variables (HAINING, 2005; CRESSIE, 1993; SANTOS and FERREIRA, 2003).

2.3 Selection for estimated tons of cane per hectare

For scenarios modeled by linear discriminant analysis, selection of the best families was also performed, selecting those with higher estimated tons of cane per hectare (TCH) than the overall mean.

The variable TCH was estimated by the following expression (FERREIRA et al., 2007):

$$TCHe = d \times \pi \times NS \times SH \times \left(\frac{SD}{2}\right)^2 \times \frac{1}{100 \times ps} \tag{1}$$

where:

d is the specific stalk density; Chang and Milligan (1992) proposed a value d of 1000 kg m^{-3};

ps is the plot size in m^2; here ps =18 m^2.

2.4 Selection by linear discriminant analysis

For the classification of selected and unselected families by discriminant analysis, two sets are needed, one to estimate and the other to test the functions.

These sets are determined by the researcher. In this study, the set used to estimate the discriminant functions was taken from experiment 1, scenario 1, and from experiment

1 plus the simulated values for 500, 750, 1000 and 2000 families in the respective scenarios 2,3,4, and 5. These research scenarios involved data simulations to enhance the estimation model (scenarios 2, 3, 4 and 5, according to the increasing number of simulated families) or involved only the original data (without simulation). For all scenarios, the test set for the functions was taken from the experiments 2, 3, 4 and 5, resulting in a total of 88 families (Table 1).

Consider $\prod_i$ the groups and $\boldsymbol{\mu}_\mathbf{i}$ and $\boldsymbol{\Sigma}_\mathbf{i}$ the multivariate means vectors and covariance matrix, respectively, of these groups, with $i = 1, 2$, since there are two classes: selected and unselected families.

It can be shown that the linear function that produces maximum separation between these populations or groups is called Fisher's linear discriminant function (FERREIRA, 2011) is written as:

$$D(\mathbf{X}) = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]^t \boldsymbol{\Sigma}^{-1}\mathbf{X}.$$

Assuming $m = \frac{1}{2}[D(\boldsymbol{\mu}_1) + D(\boldsymbol{\mu}_2)]$ as the midpoint between the two univariate population means $D(\boldsymbol{\mu}_1)$ and $D(\boldsymbol{\mu}_2)$, the following classification rule was established:

$$\mathbf{X} \in \prod_1 \text{if } D(\mathbf{X}) = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]^t \boldsymbol{\Sigma}^{-1}\mathbf{X} \geq m$$

$$\mathbf{X} \in \prod_2 \text{if } D(\mathbf{X}) = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]^t \boldsymbol{\Sigma}^{-1}\mathbf{X} < m.$$

In practice, the mean vectors and covariance matrices are unknown. Thus, the estimators of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ had to be obtained (FERREIRA, 2011).

Then, the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ were replaced by the respective estimators, $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ and $\mathbf{S}_c$. Thus we define the Fisher's sample linear discriminant function as:

$$D(\mathbf{x}) = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]^t \mathbf{S}_c^{-1}\mathbf{x}$$

Taking $m = \frac{1}{2}(D(\bar{\mathbf{x}}_1) + D(\bar{\mathbf{x}}_2))$ as the midpoint between the two univariate sample means, a sample-based classification rule can be established. This rule was applied for the classification in selected ($\prod_1$) and unselected families ($\prod_2$) ,by:

$$\mathbf{X} \in \prod_1 \text{if } D(\mathbf{X}) = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]^t \mathbf{S}_c^{-1}\mathbf{X} \geq m$$

$$\mathbf{X} \in \prod_2 \text{if } D(\mathbf{X}) = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]^t \mathbf{S}_c^{-1}\mathbf{X} < m,$$

where $\mathbf{X}$ is a vector of observations containing NS, SD and SH family means to be classified.

Discriminant analysis was performed using the "lda" function of the package MASS (VENABLES and RIPLEY, 2002) in R computing environment (R Core Team, 2013).

2.5 Evaluation and comparison of the methods

To evaluate the methods - discriminant analysis and TCHe - the apparent error rate (AER) was used. This rate, according to Cruz and Carneiro (2006), can be written as follows:

$$AER = \frac{1}{N}\sum_{j=1}^{g} e_j$$

where:

e_j is the number of families in population Π_j that were misclassified;

g the number of groups or populations; and

N the total number of families.

3 Results and discussion

As shown (Table 1), the results of discriminant analysis were good. The higher apparent error rate by discriminant analysis (0.19318), in scenario 1, without simulation, can be considered a low value, since it corresponds to a concordance rate of 80.682% with the family selection method for TCHr. By modeling with discriminant analysis, concordance rates reached 85.227% in scenarios 4 and 5, with simulated values for 1000 and 2000families, respectively.

Table 1 - Apparent error rate for discriminant analysis (AER-AD) and for estimated tons of cane per hectare (TCHe) in five scenarios

Scenarios	Simulation	Estimation exp.	Test exp	AER-SD	AER-TCHe
1	0	1	2,3,4,5	0.19318	0.17045
2	500	1	2,3,4,5	0.17045	0.17045
3	750	1	2,3,4,5	0.15909	0.17045
4	1000	1	2,3,4,5	0.14773	0.17045
5	2000	1	2,3,4,5	0.14773	0.17045

*Estimation exp. = Experiment to estimate the discriminant functions; Test exp.= Experiments to test the discriminant functions.

The great advantage of using discriminant functions to classify selected or unselected families is that only a small part of the material would have to be weighed. In this study, when weighing the harvest of only one of the five experiments (Experiment 1), the generalization for the other four experiments (Experiments 2, 3, 4 and 5) was excellent. This clearly shows that this strategy could considerably reduce the field work, optimizing the selection process.

The results of family selection for estimated tons of cane per hectare (TCH) were also good. The apparent error rate (0.17045) represents a concordance rate of 82.955% with selection for TCHr (Table 1). It is worth mentioning that the simulation does not interfere with TCHe selection, since original data were used.

Comparatively, the apparent error rate for discriminant analysis (AER-AD) was lower than or equal to the apparent error rate for estimated tons of cane per hectare in scenarios with simulation (scenarios 2, 3, 4 and 5) and highest in scenario 1 with simulation (Table 1). This result is best seen in Figure 1.

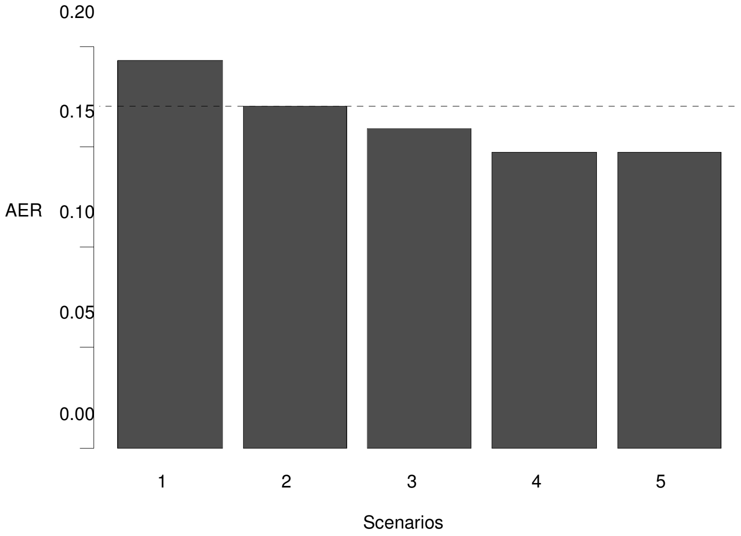

Figure 1 - Apparent error rate (AER) for discriminant analysis in five scenarios (1: No simulation, 2: simulation of 500 families, 3: simulation of 750 families; 4: simulation of 1000 families, 5: simulation of 2000 families) and the apparent error rate for estimated tons of cane per hectare indicated by the dashed line.

In addition, a decrease in the apparent error rate can be observed with increasing number of simulated families, stabilizing in scenario 4 when SN, SD, SH, and TCHr values were simulated for 1000 families (Figure 1).

Simulation is important because the original data set to estimate the discriminant functions is very small, affecting the discrimination ability of the functions. In studies of adaptability and stability in alfalfa genotypes, Barroso et al. (2013) also applied simulation to compose a data set to estimate discriminant functions, with good results.

The question arises whether the results would be the same if another experiment, different from 1, were used for simulation and to constitute the data set to estimate the functions. However, the differences between the variables NS, SD, SH, and TCHr in the five experiments were considerable (Figure 2).

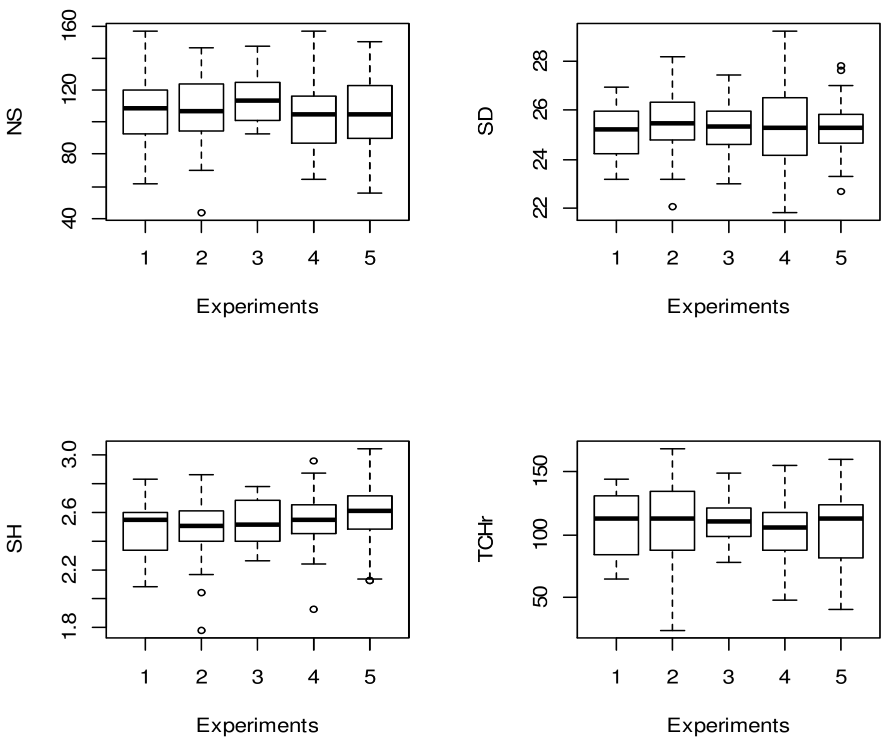

Figure 2 - Box plot for the variables number of stalks (NS), stalk diameter (SD), in centimeters, stalk height (SH), in meters and real tons of cane per hectare (TCHr).

In fact, the results of a replicate of the same analysis for the best scenario (Scenario 4), based on each of the five experiments separately for the simulation and to compose the set for estimation of the functions, were similar (Table 2).

Despite the good results obtained by modeling with discriminant analysis, further studies are needed to evaluate the real effectiveness of the technique. These studies would clarify, for example, if modeling with discriminant analysis is more efficient than other possible classification techniques, such as artificial neural networks and logistic regression.

Another question that arises is whether the performance of the discriminant functions would still be superior at lower selection rates. In this study, assuming normal distribution, the selection rate was approximately 50%, since the TCHr of the selected families was generally above the overall experimental mean.

Table 2 - Apparent error rate for discriminant analysis (AER-AD) and for estimated tons of cane per hectare in scenario 4, using different situations for estimation and testing of the discriminant functions.

Scenario	Simulation	Estimation exp.	Test exp.	AER-SD	AER-TCHe
4	1000	1	2,3,4,5	0.14772	0.17045
4	1000	2	1,3,4,5	0.125	0.18182
4	1000	3	1,2,4,5	0.18182	0.14773
4	1000	4	1,2,3,5	0.14773	0.15909
4	1000	5	1,2,3,4	0.13636	0.15909

*Estimation exp. = Experiment to estimate the discriminant functions; Test exp.= Experiments to test the discriminant functions.

Conclusions

The linear discriminant analysis based on the indirect traits NS, SD and SH presented a high concordance with the selection based on TCHr.

Selection of families based on linear discriminant analysis has the advantage of minimize the field labor because only part of the family plots should be weigh.

On the scenarios with simulation the family selection based on linear discriminant analysis was better than the selection based on TCHe.

At least 1,000 simulations to augment the training population are recommended to minimize the apparent error rate of selection when using the linear discriminant analysis.

Acknowledgements

The authors are indebted to Brazilian Council for Scientific and Technological Development (CNPq), Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) and Research Foundation of the State of Minas Gerais (FAPEMIG for funding.

MOREIRA, E. F. A.; PETERNELLI, L. A.; Seleção precoce entre famílias de cana-de-açúcar via classificação por análise discriminante linear. *Rev. Bras. Biom.*, São Paulo, v.33, n.4, p.484-493, 2015.

- Resumo: Um dos grandes desafios nos programas de melhoramento genético de cana-de-açúcar é a seleção eficiente de genótipos nas fases iniciais. Uma seleção eficiente nestas fases é de suma importância para os objetivos do programa, uma vez que, devido á particularidade da cana-de-açúcar, os materiais selecionados na fase inicial serão avaliados nas etapas posteriores. O objetivo deste trabalho é comparar a seleção via análise discriminante linear e a seleção de famílias usando a variável tonelada de cana por hectare estimada (TCHe) com base nos caracteres indiretos número de colmos, diâmetro de colmos e altura de colmos como alternativas para seleção de famílias promissoras em cana-de-açúcar. Também foi considerado o uso de simulação para aumento do conjunto de treinamento previamente a análise. As análises foram realizadas em cinco diferentes cenários, definidos em função do número de valores simulados: sem simulação, com simulação de valores para 500, 750, 1000 e 2000 famílias. Para comparação e avaliação dos métodos empregados foi utilizada a taxa de erro aparente (TEA). A análise discriminante linear apresenta alta concordância com a seleção via TCHr podendo ser utilização para seleção precoce entre famílias em cana-de-açúcar. Nos cenários onde houve simulação, a análise discriminante linear tem desempenho superior a seleção via TCHe. A taxa de erro aparente é minimizada quando pelo menos 1.000 famílias são utilizadas para o aumento da população de treinamento.

- PALAVRAS-CHAVE: Simulação; melhoramento de plantas; Saccharum spp.

References

BARBOSA, M. H. P.; SILVEIRA, L. C. I. Melhoramento Genético e Recomendação de Cultivares. In: SANTOS, F.; BORÉM, A.; CALDAS, C. (Ed.). *Cana-de-açúcar: Bioenergia, Açúcar e Álcool - Tecnologias e Perspectivas*. Viçosa, MG: Suprema, 2010. p. 313-331.

BARROSO, L. M. A.; NASCIMENTO, M**.**; NASCIMENTO, A. C. C.; SILVA, F. F.; FERREIRA, R. P. Uso do método de Eberhart e Russel como informação a priori para aplicação de redes neurais artificiais e análise discriminante visando a classificação de genótipos de alfafa quanto à adaptabilidade e estabilidade fenotípica. *Revista Brasileira de Biometria*, v. 31, p. 176-188, 2013.

BRASIL. Companhia Nacional de Abastecimento. Acompanhamento da Safra Brasileira. Safra 2014/2015. 2° Levantamento da Cana-de-Açúcar 2014 Brasília: Conab, 2014. 19p. Disponível em:< http://www.conab.gov.br/OlalaCMS/uploads/arquivos/14_08_28_08_52_35_boletim_cana_portugues_-_2o_lev_-_2014-15.pdf>. Acesso em: 2 dez. 2014.

CHANG, Y.S.; MILLIGAN S.B. Estimating the potential of sugarcane families to produce elite genotypes using univariate cross prediction methods. *Theoretical and Applied Genetics*, Berlin, v.84, p.662-671, 1992.

CRESSIE, N. A. C. *Statistics for Spatial data*. John Wiley& Sons, Inc.1993. 900p. .

FERREIRA, D. F. *Estatística Multivariada*. 2.ed, Lavras: Ed. UFLA, 2011. 675p.

FERREIRA, F. M.; BARROS, W. S.; SILVA, F. L.; BARBOSA, M. H. P.; CRUZ, C. D.; BASTOS, I. T. Relações fenotípicas e genotípicas entre componentes de produção em cana-de-açúcar. *Bragantia*, Campinas, v.66, n.4, p.527-533, 2007.

HAINING, R. *Spatial data analysis – theory and practice.* Cambridge University Press, 2005. 432p

KHATTREE, R.; NAIK, D. N. *Multivariate Data Reduction and Discrimination with SAS Software,*1.ed, Cary: SAS Institute Inc. 2000. 558p.

NOGUEIRA, A. P. O.; SEDIYAMA, T.; CRUZ, C. D.; REIS, M. S.; PEREIRA, D. G.;JANGARELLI, M. Novas características para diferenciação de cultivares de soja pela análise discriminante.*Ciência Rural*, Santa Maria, v.38, n.9, p.2427-2433, 2008.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.r-project.org).

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

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

SANTOS, A. C.; FERREIRA, D.F. Definição do tamanho amostral usando simulação Monte Carlo para o teste de normalidade baseado em assimetria e curtose. II. Abordagem Multivariada. *Ciência Agrotécnica*, v.27, n.1, p.62-69, 2003.

STRINGER, J. K.; COX, M. C.; ATKIN, F. C.; WEI, X.; HOGARTH. Family Selection Improves the Efficiency and Effectiveness of Selecting Original Seedlings and Parents. *Sugar Tech*, Kunraghat, v.13, n.1, p.36–41, 2011.

SUDRÉ, C. P.; CRUZ, C. D.; RODRIGUES, R.; RIVA, E. M.; AMARAL JÚNIOR, A.T.; SILVA, D. J. H., PEREIRA, T. N. S. Variáveis multicategóricas na determinação da divergência genética entre acessos de pimenta e pimentão. *Horticultura Brasileira*, Campinas, v.24, n.1, p.88-93, 2006.

VENABLES W. N.; RIPLEY B. D. *Modern applied statistics with **S*****.** New York: Springer, 4.ed, 2002. 493p.

WACLAWOVSKY, A. J.; SATO, P. M.; LEMBKE, C. G.; MOORE, P. H.; SOUZA, G. M. Sugarcane for bioenergy production: an assessment of yield and regulation of sucrose content. *Plant Biotechnology Journal*, v.8, n.3, p.263-276, 2010.

Recebido em 24.03.2015

Aprovado após revisão em 27.08.2015