DEICIANA PAGANO ESPÓSITO

# ANÁLISE DE TRILHA EM DADOS DE PRODUÇÃO E TECNOLÓGICOS DA CANA-DE-AÇÚCAR

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
2010

DEICIANA PAGANO ESPÓSITO

# ANÁLISE DE TRILHA EM DADOS DE PRODUÇÃO E TECNOLÓGICOS DA CANA-DE-AÇÚCAR

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: 04 de fevereiro de 2010.

_____ _____
Prof. Márcio Henrique Pereira Barbosa
(Co-Orientador)

Prof. Cosme Damião Cruz
(Co-Orientador)

_____ _____
Prof. Antônio Policarpo Souza Carneiro

Prof. Paulo Roberto Cecon

_____
Prof. Luiz Alexandre Peternelli
(Orientador)

A Deus, por guiar meu caminho.

Aos meus pais, Ademar e Marília, que com muito amor, me mostraram os diferentes caminhos da vida, dando-me apoio e liberdade para seguir aquele que escolhi, e por estarem sempre ao meu lado.

Aos meus irmãos, Rafael e Ademar Júnior, pela amizade e confiança.

OFEREÇO

Ao meu namorado Marcello, pelo carinho, apoio nos momentos difíceis, e por fazer parte da minha vida.

DEDICO

## AGRADECIMENTOS

A Deus, pela minha vida, por me guiar, fortalecer, iluminar meu caminho, ajudando-me a vencer os obstáculos, e pela proteção em todos os momentos.

À minha família, Ademar, Marília, Ademar Júnior, Rafael, Walkíria, Carolyne e Ana Rafaela, pelo apoio, incentivo e confiança.

Ao meu namorado Marcello, pela compreensão, pelo estímulo, amor e carinho.

Ao meu orientador Luiz Alexandre Peternelli, pelo valioso ensino, incentivo, confiança, amizade e pela convivência que com certeza foi fundamental na minha formação acadêmica e pessoal.

Aos professores Cosme Damião Cruz e Márcio Henrique Pereira Barbosa**,** por todas as sugestões, ensinamentos e disponibilidade.

Aos professores do Curso de Mestrado em Estatística Aplicada e Biometria, pelos ensinamentos, atenção e pelo caráter humano transmitido aos seus alunos e orientados.

Ao secretário da Pós-Graduação Altino, pelo empenho em sempre ajudar e pelas mensagens de amizade e incentivo.

Aos colegas do Curso de Estatística Aplicada e Biometria, pela valiosa amizade e companheirismo.

À banca, composta por Márcio Henrique Pereira Barbosa, Cosme Damião Cruz, Antônio Policarpo Souza Carneiro e Paulo Roberto Cecon, que

aceitaram o convite que lhes foi feito e, dessa forma, colaboraram para conclusão deste trabalho.

À Universidade Federal de Viçosa e ao Departamento de Estatística, pela oportunidade de realização do curso.

À Secretaria de Estado de Educação de Minas Gerais, pela oportunidade concedida para a realização deste curso.

À Escola Estadual do Ribeirão de São Domingos pelo apoio.

A todos os meus parentes e amigos, sempre presentes na minha vida pessoal e profissional.

A todos que, direta ou indiretamente, colaboraram para a realização deste trabalho.

**SUMÁRIO**

## RESUMO

ESPÓSITO, Deiciana Pagano, M. Sc., Universidade Federal de Viçosa, fevereiro de 2010. **Análise de trilha em dados de produção e tecnológicos da cana-de-açúcar**. Orientador: Luiz Alexandre Peternelli. Co-orientadores: Márcio Henrique Pereira Barbosa e Cosme Damião Cruz.

Com o objetivo de quantificar os efeitos diretos e indiretos, por meio da análise de trilha, utilizando valores fenotípicos e genotípicos dos componentes de produção – número de colmos por parcela, diâmetro médio de colmos e comprimento médio de colmos – sobre produtividade de colmos por hectare em cana-de-açúcar, foram obtidos dados de dois experimentos nas fases de cana-planta e cana-soca, em etapa inicial de seleção do programa de melhoramento da cana-de-açúcar no estado de Minas Gerais. Foram avaliados, ao nível de parcela, os caracteres tonelada de colmos por hectare (*TCH*), como variável principal, e seus componentes de produção, número de colmos (*NC*), diâmetro médio de colmos (*DC*) e comprimento médio de colmos (*CC*), como variáveis explicativas. Os coeficientes de determinação foram elevados em todas as análises de trilha, indicando que os componentes avaliados explicam grande parte da variação existente na produção de colmos. Pela análise dos efeitos diretos fenotípicos e genotípicos, *NC* foi a variável que melhor se correlacionou com *TCH*, em ambos os experimentos e estágios, demonstrando a possibilidade de obtenção de ganhos significativos por meio da seleção indireta para *TCH* via *NC*. A avaliação das relações de causa e efeito entre os componentes de produção em cana-de-açúcar possibilitou verificar que houve variação entre os experimentos, o que provavelmente se deve à origem diferenciada das famílias avaliadas. Como na técnica de análise de trilha os parâmetros são estimados a partir de matrizes de correlações que podem ser mal condicionadas por efeito de multicolinearidade entre as variáveis envolvidas, foram avaliados dados em cana-soca, obtidos do programa de melhoramento da cana-de-açúcar da Universidade Federal de Viçosa, para comparar o método baseado na regressão em crista e a exclusão de

variáveis por componentes principais para a estimação dos coeficientes de trilha em presença de multicolinearidade. Foram amostradas dez plantas por parcela para realização das análises das variáveis explicativas Brix (teor de sólidos solúveis), Pol (teor de sacarose aparente), pH (indica o grau de acidez), AR (açúcares redutores), ART (açúcares totais recuperáveis), Cu (cobre), Al (alumínio), Mg (magnésio), Ca (cálcio), K (potássio), Ácido aconítico, Compostos fenólicos, e da variável principal Cor ICUMSA. A matriz de correlação obtida dos dados foi submetida a diferentes métodos para diagnóstico de multicolinearidade. Sob multicolinearidade severa, os métodos baseados na regressão em crista e em componentes principais apresentaram resultados semelhantes na estimação dos coeficientes de trilha, proporcionando sensível redução na magnitude dos fatores de inflação da variância associados aos efeitos diretos e indiretos da análise de trilha. Assim, foi possível identificar neste estudo, os caracteres alumínio (Al), potássio (K) e Compostos fenólicos como aqueles que melhor explicam a Cor do caldo. Contudo, os demais caracteres devem ser levados em consideração devido a elevada correlação existente e a baixa magnitude do efeito direto, evidenciando a necessidade de seleção simultânea de caracteres, com ênfase também nos caracteres cujos efeitos indiretos são significativos. Para fins de melhoramento, a seleção indireta para Cor do caldo, por meio de índice de seleção envolvendo as variáveis Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos e Ácido aconítico é recomendável.

## ABSTRACT

ESPÓSITO, Deiciana Pagano, M. Sc., Universidade Federal de Viçosa, February, 2010. **Path analysis for yield components and technological data of sugarcane**. Advisor: Luiz Alexandre Peternelli. Co-Advisors: Márcio Henrique Pereira Barbosa and Cosme Damião Cruz.

In order to quantify the direct and indirect effect through path analysis using phenotypic and genotypic values of yield components – number of stalks per plot, average diameter of stalks and their average length – affecting tons of cane per hectare, data from two experiments of sugarcane were considered. Regarding the productivity of cane per hectare data from two different experiments were obtained from cane plants and ratoon canes, in the beginning of the selection program of sugarcane improvement in the state of Minas Gerais. The following characteristics were evaluated at plot level: the tons of cane per hectare (TCH), as the main variable ,and its yield components, number of stalks (NS), mean diameter of stalks (DS) and average length of stalks (LS) as explicative variables. The coefficient of determination were high in all path analyses, which, in turn, indicates that the evaluated components explain, considerably, the variation in TCH. Through the analysis of the direct fenotypic and genotypic effects, NS was the variable that best correlated to TCH in both experiments and stages showing a possibility of obtaining significant gain through indirect selection to TCH by NS. The evaluation of the cause and effect relations among the production components of sugarcane helped to verify the variation across the experiments, which is probably related to the different origins of the families evaluated. In the trail analysis, the parameters are estimated from matrix correlations that may be ill-conditioned by the multicollinearity effect among the involved variables. Due to this fact, the data were evaluated by using ratoon canes obtained from the program of sugarcane improvement at the Federal University of Viçosa in order to compare the method based on ridge regression and the exclusion of variables for main components to estimate

the path coefficients under the presence of multicollinearity. Ten plants per plot were used to carry out the analyses on explaining variables Brix (percentage soluble solids), Pol, pH (potential of Hydrogen), RS (reduction sugar), TRS (Total reduction sugar)**,** Cu (copper), Al (aluminum), Mg (magnesium), Ca (Calcium), K (Potassium), aconitic acid, phenolic compounds and the main variable sugarcane juice color (ICUMSA color)**.** The matrix containing correlation obtained from the data were submitted to different methods to have the multicollinearity diagnostic. Under severe multicollinearity, the methods based on ridge regression and in main components presented similar results in the estimation of the path coefficients, causing sensitive reduction in the magnitude of the variance inflation factor associated with the direct and indirect effects of the path analysis. Therefore, in this study, it was possible to identify the variables Al, K and phenolic compounds as the ones that explain the sugarcane juice color. However, the other characters must be taken into account due to their great correlation and low magnitude of the direct effect, making evident the necessity of simultaneous selections of characters, with emphasis on characters that have significant indirect effect. For purposes of improvement, the indirect selection for the ICUMSA color through index selection involving variables such as Brix, Pol, RS, TRS, pH, Cu, Al, Mg, Ca, K, phenolic compounds and aconitic acid is recommended.

# 1. INTRODUÇÃO GERAL

Atualmente a cana-de-açúcar é uma das principais culturas agrícolas do país. Na economia brasileira, ela representa um papel especialmente importante para a produção de açúcar e álcool principalmente, com área cultivada de 8,5 milhões de hectares, de acordo com dados relativos ao terceiro levantamento de safra feito pela Companhia Nacional de Abastecimento – CONAB, 2009 (Goes et al., 2009).

Nos últimos 30 anos, a cultura da cana-de-açúcar incorporou tecnologias capazes de aumentar os níveis de produtividade. O melhoramento genético foi a ferramenta que garantiu e continuará garantindo a sustentabilidade dessa cultura, tornando-se resistente a pragas e doenças, reduzindo custos, aumentando a eficiência e a produtividade, com a disponibilização de novas variedades de plantas adaptáveis às condições de cada região (Goes et al., 2009).

O conhecimento das correlações entre os caracteres tem grande importância em programas de melhoramento, principalmente quando a seleção de um caráter desejável apresenta dificuldades e/ou, problemas de medição e identificação, por se tratar de um caráter de baixa herdabilidade (Cruz, 2006).

As relações existentes entre os caracteres são, em geral, avaliadas por meio das correlações genotípicas, fenotípicas e de ambiente. Tais estudos têm sido amplamente utilizados para medir o grau de associação entre caracteres, em diversas culturas, como na moranga, por Amaral Júnior et al. (1994); na soja, por Peluzio et al. (1998); no pimentão, por Miranda et al.,(1988). Nesses estudos, a correlações genotípicas se mostraram, na maioria das vezes, superiores às fenotípicas, indicando herdabilidade elevada dos caracteres e evidenciando uma contribuição satisfatória dos fatores genéticos nas correlações em estudo. Para Carvalho et al. (2001), citado por Benin et al. (2003), a seleção de caracteres de alta herdabilidade, fácil aferição e identificação, e que evidencie alta correlação com o caráter

desejado, possibilita ao melhorista obter maior progresso, em menor espaço de tempo.

Entretanto, essas correlações não representam uma medida de causa e efeito, e não provêem informações a respeito dos efeitos diretos e indiretos de um grupo de caracteres em relação a um determinado caráter considerado de maior importância. Segundo Ferreira et al. (2007), a interpretação direta das suas magnitudes pode resultar em equívocos na estratégia de seleção, pois uma alta correlação entre duas variáveis pode ser resultado do(s) efeito(s) de outra(s) sobre elas.

Dessa forma, informações mais úteis e indispensáveis no melhoramento podem ser obtidas por meio da análise de trilha (path analysis), desenvolvida pelo geneticista Sewall Wright em 1918-1921, segundo Johnson e Wichern (1992). Essa técnica tem sido utilizada no estudo dos efeitos diretos e indiretos de um conjunto de variáveis sobre um determinado caráter, considerado como principal, caracterizado por grande complexidade, porém, de considerável importância econômica.

A decomposição da correlação depende do conjunto de caracteres estudados, os quais, normalmente são avaliados pelo conhecimento prévio do pesquisador de sua importância e de possíveis inter-relações expressas em “diagramas de trilha” (Cruz et al., 2004). A construção gráfica de esquema causal possibilita a obtenção de um conjunto de equações simultâneas.

O método da análise de trilha tem sido eficiente para revelar a verdadeira natureza das inter-relações de causa e efeito entre um caráter de importância econômica e seus componentes. O emprego desta técnica para estudo do relacionamento entre caracteres em plantas teve início com Dewey e Lu (1959), sendo atualmente aplicada em diversas culturas, inclusive na cana-de-açúcar.

Com a finalidade de se obter subsídios para o melhoramento da cana-de-açúcar, tem-se dado ênfase ao estudo das correlações entre caracteres determinantes da produção e das técnicas de identificação de genótipos (ou indivíduos) promissores com base em características indiretas, pois embora o

caráter produção seja o principal objetivo no melhoramento da cana-de-açúcar, é preciso considerar outras variáveis na estratégia de seleção.

Os componentes associados à produtividade da cana-de-açúcar apresentam herança complexa e provavelmente controlados por genes de efeitos pleiotrópicos com outros caracteres, justificando a identificação e a análise de outras variáveis de controle genético menos complexo associadas a essas variáveis principais.

O emprego da análise de trilha nos programas de melhoramento da cana-de-açúcar é de grande valia nas suas etapas iniciais. Esta metodologia visa apontar as características mais adequadas para que seja feita uma seleção indireta dos genótipos mais produtivos, uma vez que quantificar a produção destes genótipos é um trabalho bastante demorado devido ao grande número de genótipos avaliados nestas etapas.

As relações entre variáveis importantes para a cultura da cana-de-açúcar têm sido objeto de estudos por meio da análise de trilha. Kang et al. (1983) demonstraram a importância do desdobramento dos coeficientes de correlação genotípica envolvendo produção de cana e seus componentes, via análise de trilha. Ferreira et al. (2007) utilizaram a análise de trilha para a cultura da cana-de-açúcar, com o objetivo de quantificar os efeitos diretos e indiretos de componentes de produção sobre as variáveis tonelada de cana por hectare e massa média de colmos.

Todavia, a técnica de análise de trilha pode apresentar dificuldades, pois seus parâmetros são estimados a partir de matrizes de correlações ou covariâncias fenotípicas ou genotípicas, as quais podem ser mal condicionadas por efeito de multicolinearidade entre as variáveis envolvidas.

Em presença de multicolinearidade, as variâncias associadas aos estimadores dos coeficientes de trilha que medem efeitos diretos de variáveis explicativas sobre uma principal, podem atingir valores demasiadamente elevados, sendo evidência de serem as estimativas pouco confiáveis, o que pode levar à interpretação equivocada ou sem nenhuma coerência com o fenômeno biológico estudado (Cruz e Carneiro, 2006).

Diagnósticos de multicolinearidade devem ser feitos de forma a viabilizar certos estudos, como a análise de trilha. Várias técnicas têm sido propostas para diagnosticar a presença de multicolinearidade, sendo características desejáveis de um procedimento de diagnóstico aquelas que refletem diretamente o grau do problema de multicolinearidade e provêm informações úteis na determinação de quais variáveis estão envolvidas (Montgomery e Peck, 1992).

Os métodos alternativos à estimação de mínimos quadrados, para contornar os efeitos da multicolinearidade e aumentar a estabilidade dos coeficientes de regressão, fornecem estimadores tendenciosos, mas, conforme Gunst e Mason (1977) apresentam, em geral, melhor desempenho quando comparados aos estimadores de mínimos quadrados.

O método de regressão em crista ou em cumeeira (Cruz e Carneiro, 2006), originalmente proposto por Hoerl e Kennard (1970a,b), e a regressão em componentes principais (Montgomery e Peck,1992), são procedimentos alternativos propostos para combater os problemas induzidos pela multicolinearidade. O método de regressão em crista se baseia em obter estimativas de coeficientes de regressão a partir de uma versão ligeiramente modificada das equações normais (Cruz e Carneiro, 2006). No segundo procedimento, os componentes principais correspondentes aos autovalores próximos de zero são removidos da análise e o método dos mínimos quadrados é aplicado aos componentes restantes.

Em virtude do estudo das relações entre variáveis importantes para a cultura da cana-de-açúcar, via análise de trilha, o presente trabalho tem por objetivo quantificar os efeitos diretos e indiretos, por meio da análise de trilha, utilizando valores fenotípicos e genotípicos dos componentes de produção – número de colmos por parcela, diâmetro médio de colmos e comprimento médio de colmos – sobre a produtividade de colmos por hectare em cana-de-açúcar, nas fases de cana-planta e cana-soca em etapa inicial de seleção do programa de melhoramento da cana-de-açúcar no estado de Minas Gerais, favorecendo, assim, o processo de seleção indireta de genótipos mais produtivos. Além disso, em virtude do dano que a multicolinearidade pode

causar ao processo de estimação dos coeficientes de trilha, o presente trabalho visa também, comparar dois métodos alternativos de contornar os efeitos adversos da multicolinearidade na estimação dos coeficientes de trilha, determinando, através da análise de trilha, o efeito dos principais constituintes orgânicos e inorgânicos sobre a cor do caldo em cultivares de cana-de-açúcar.

# 2. REFERENCIAL TEÓRICO

## 2.1. Correlações genotípica, fenotípica e de ambiente

O grau de associação linear entre duas variáveis quaisquer pode ser definido pela correlação de Pearson (Ferreira, 2007). A correlação simples permite avaliar a magnitude e o sentido das relações entre dois caracteres, sendo de grande utilidade no melhoramento, por permitir avaliar a viabilidade da prática da seleção indireta, que, em alguns casos, pode levar a progressos mais rápidos que a seleção do caráter desejado (Cruz, 2006).

Segundo Falconer (1987), em estudos genéticos, é necessário distinguir duas causas de correlação entre características: a genética, resultante de ligação gênica (causa de correlação transitória) ou do pleiotropismo (propriedade pela qual um gene afeta duas ou mais características), e a de ambiente. O ambiente é uma causa de correlação quando duas características são influenciadas pelas mesmas diferenças de condições de ambiente. A associação entre duas características que pode ser observada diretamente é a correlação de valores fenotípicos, ou a correlação fenotípica.

### 2.1.1. Estimação dos coeficientes de correlação

O estimador do coeficiente de correlação entre duas variáveis aleatórias $X$ e $Y$ é dado pela expressão

$$r_{XY} = \frac{\hat{Cov}(X,Y)}{\sqrt{\hat{V}(X)\cdot\hat{V}(Y)}} = \frac{\frac{SPD_{XY}}{n-1}}{\sqrt{\frac{SQD_X}{n-1}\cdot\frac{SQD_Y}{n-1}}} = \frac{SPD_{XY}}{\sqrt{SQD_X \cdot SQD_Y}},$$

em que

$$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n},$$

$$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} \qquad \text{e} \qquad SQD_Y = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}.$$

As variáveis *X* e *Y* seguem distribuição normal bivariada conforme demonstrado por Montgomery e Peck (1992).

Pode ser provado que a correlação está sempre no intervalo [-1, 1], como apresentado por Bain e Engelhardt, 1992; Casella e Berger, 2002.

De acordo com Cruz et al. (2004), para estimação dos coeficientes de correlação genotípica, fenotípica e de ambiente entre dois caracteres *X* e *Y*, recomenda-se as análises individuais, segundo um modelo estatístico apropriado, e a análise da soma dos valores de *X* e *Y*, de tal forma que os produtos médios (covariâncias), associados a cada fonte de variação, possam ser estimados por meio de

$$Cov(X,Y) = \frac{V(X+Y) - V(X) - V(Y)}{2}$$

Os componentes de covariância podem ser estimados por meio da esperança do produto médio das fontes de variações, que são obtidas de maneira equivalente às esperanças dos respectivos quadrados médios da análise de variância, sendo necessário apenas substituir a expressão de variância pela covariância.

Considerando os caracteres $X_{ij}$ e $Y_{ij}$, medidos em *g* genótipos ou tratamentos (i = 1, 2, ..., *g*), avaliados em blocos ao acaso com *r* repetições (j = 1, 2, ..., *r*), tem-se o esquema da análise de variância apresentado na Tabela 1.

**Tabela 1.** Esquema da análise de variância dos caracteres *X*, *Y* e da soma *X* + *Y*, para o experimento em blocos casualizados.

| | | *QM* | | | |
|---|---|---|---|---|---|
| FV | GL | *X* | *Y* | *X+Y* | *E(QM)* |
| Blocos | r –1 | | | | |
| Tratamentos | g – 1 | $QMT_x$ | $QMT_y$ | $QMT_{x+y}$ | $\sigma^2 + r\sigma_g^2$ |
| Resíduo | (r –1) (g – 1) | $QMR_x$ | $QMR_y$ | $QMR_{x+y}$ | $\sigma^2$ |

Os produtos médios associados a tratamentos e resíduo são obtidos, respectivamente, por meio das expressões

$$PMT_{xy} = (QMT_{x+y} - QMT_x - QMT_y)/2$$

e

$$PMR_{xy} = (QMR_{x+y} - QMR_x - QMR_y)/2 .$$

O esquema da análise, incluindo os produtos médios e suas respectivas esperanças matemáticas, é apresentado na Tabela 2.

**Tabela 2.** Esquema da análise com os produtos médios e suas respectivas esperanças matemáticas, para o experimento em blocos casualizados.

| FV | GL | *PM* | *E(PM)* |
|---|---|---|---|
| Blocos | r –1 | | |
| Tratamentos | g – 1 | $PMT_{xy}$ | $\sigma_{xy} + r\sigma_{gxy}$ |
| Resíduo | (r –1) (g – 1) | $PMR_{xy}$ | $\sigma_{xy}$ |

Com base nos resultados das análises apresentadas nas Tabelas 1 e 2, os coeficientes de correlação são estimados por meio das expressões descritas a seguir:

i) Correlação fenotípica

$$r_f = \frac{PMT_{xy}}{\sqrt{QMT_x QMT_y}}$$

ii) Correlação de ambiente

$$r_a = \frac{PMR_{xy}}{\sqrt{QMR_x QMR_y}}$$

iii) Correlação genotípica

$$r_g = \frac{\hat{\sigma}_{gxy}}{\sqrt{\hat{\sigma}^2_{gx}\hat{\sigma}^2_{gy}}}$$

Sendo

$$\hat{\sigma}_{gxy} = \frac{PMT_{xy} - PMR_{xy}}{r};$$

$$\hat{\sigma}^2_{gx} = \frac{QMT_x - QMR_x}{r} \quad \text{e}$$

$$\hat{\sigma}^2_{gy} = \frac{QMT_y - QMR_y}{r}.$$

em que

$\hat{\sigma}_{gxy}$: estimador da covariância genotípica entre os caracteres *X* e *Y*;

$\hat{\sigma}^2_{gx}$ e $\hat{\sigma}^2_{gy}$: estimador das variâncias genotípicas dos caracteres *X* e *Y,* respectivamente.

A herdabilidade ($h^2_x$) pode ser estimada por meio de:

$$h^2_x = \frac{\hat{\sigma}^2_{gx}}{\hat{\sigma}^2_{fx}},$$

em que

$\hat{\sigma}^2_{fx}$: estimador da variância fenotípica do caráter *X*.

Como

$$\hat{\sigma}^2_{fx} = \frac{QMT_x}{r},$$

tem-se que

$$h^2_x = \frac{\hat{\sigma}^2_{gx}}{\frac{QMT_x}{r}}$$

$$h^2_x = \frac{r\hat{\sigma}^2_{gx}}{QMT_x}$$

$$QMT_x = \frac{r\hat{\sigma}^2_{gx}}{h^2_x}.$$

Por analogia,

$$h^2_y = \frac{\hat{\sigma}^2_{gy}}{\hat{\sigma}^2_{fy}} \quad \text{e}$$

$$QMT_y = \frac{r\hat{\sigma}^2_{gy}}{h^2_y}.$$

O complemento aritmético da herdabilidade é dado por:

$$e^2_x = 1 - h^2_x.$$

Sendo,

$$e^2_x = 1 - \frac{\hat{\sigma}^2_{gx}}{\hat{\sigma}^2_{fx}}$$

$$e^2_x = \frac{\hat{\sigma}^2_{fx} - \hat{\sigma}^2_{gx}}{\hat{\sigma}^2_{fx}},$$

onde

$$\hat{\sigma}^2_{fx} = \frac{QMT_x}{r} = \frac{\sigma^2_x + \sigma^2_{gx}}{r} = \frac{\hat{\sigma}^2_x}{r} + \hat{\sigma}^2_{gx},$$

tem-se que,

$$e^2_x = \frac{\left(\frac{\hat{\sigma}^2_x}{r} + \hat{\sigma}^2_{gx}\right) - \hat{\sigma}^2_{gx}}{\hat{\sigma}^2_{fx}} = \frac{\hat{\sigma}^2_x}{r\hat{\sigma}^2_{fx}}$$

$$e^2_x = \frac{\hat{\sigma}^2_x}{QMT_x} \Rightarrow QMT_x = \frac{\hat{\sigma}^2_x}{e^2_x}.$$

Por analogia,

$$e^2_y = \frac{\hat{\sigma}^2_y}{r\hat{\sigma}^2_{fy}}$$

$$e^2_y = \frac{\hat{\sigma}^2_y}{QMT_y} \Rightarrow QMT_y = \frac{\hat{\sigma}^2_y}{e^2_y}.$$

Como

$$r_f = \frac{PMT_{xy}}{\sqrt{QMT_x QMT_y}} = \frac{r\hat{\sigma}_{gxy} + \hat{\sigma}_{xy}}{\sqrt{QMT_x QMT_y}}$$

$$= \frac{r\hat{\sigma}_{gxy}}{\sqrt{QMT_x QMT_y}} + \frac{\hat{\sigma}_{xy}}{\sqrt{QMT_x QMT_y}}$$

$$= \frac{r\hat{\sigma}_{gxy}}{\sqrt{\left(\frac{r\hat{\sigma}_{gx}^2}{h_x^2}\right)\left(\frac{r\hat{\sigma}_{gy}^2}{h_y^2}\right)}} + \frac{\hat{\sigma}_{xy}}{\sqrt{\left(\frac{\hat{\sigma}_x^2}{e_x^2}\right)\left(\frac{\hat{\sigma}_y^2}{e_y^2}\right)}}$$

$$= \frac{r\hat{\sigma}_{gxy}}{\frac{r\hat{\sigma}_{gx}\hat{\sigma}_{gy}}{h_x h_y}} + \frac{\hat{\sigma}_{xy}}{\frac{\hat{\sigma}_x\hat{\sigma}_y}{e_x e_y}}$$

$$= \frac{\hat{\sigma}_{gxy}h_x h_y}{\hat{\sigma}_{gx}\hat{\sigma}_{gy}} + \frac{\hat{\sigma}_{xy}e_x e_y}{\hat{\sigma}_x\hat{\sigma}_y},$$

constata-se que $r_f = h_x h_y r_g + e_x e_y r_a$ .

A partir da expressão $r_f = h_x h_y r_g + e_x e_y r_a$ , é possível observar alguns fatos interessantes:

a) Se $h_x^2 = 1; \quad e_x^2 = 0 \quad \Rightarrow r_f = h_y r_g \quad \Rightarrow r_g = \frac{r_f}{h_y}$.

Logo, a correlação genotípica pode ser maior que a correlação fenotípica, quando a herdabilidade é muito alta.

b) Se $h_x^2 = h_y^2 = \frac{1}{2}; \quad e_x^2 = \frac{1}{2} \quad \Rightarrow r_f = \frac{r_g + r_a}{2}$.

c) Se $h_x^2 = 0; \quad e_x^2 = 1 \quad \Rightarrow r_f = r_a$

### 2.1.2 Teste de significância para o coeficiente de correlação

De acordo com Dunn e Clark (1974), a hipótese de que o coeficiente de correlação é igual a zero; isto é,

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0 \ ,$$

pode ser testada pelo teste *t*, dado por

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Se $\rho$ é realmente zero, esta estatística tem uma distribuição *t* com n-2 graus de liberdade. O *t* calculado dos dados é comparado com $t_{\alpha/2,n-2}$ ; se

esta magnitude é maior que $t_{\alpha/2,n-2}$, conclui-se que $\rho$ é diferente de zero, isto é, a hipótese de nulidade é rejeitada.

De acordo com Vencovsky e Barriga (1992), quando a correlação é fenotípica e o coeficiente de correlação é calculado diretamente de um conjunto de n pares de dados, a significância pode ser verificada em tabela apropriada, com n-2 graus de liberdade, conforme descrito anteriormente. Porém, a consulta desta tabela não é válida quando o coeficiente de correlação é estimado a partir de componentes da variância e da covariância, como acontece no caso da correlação genética entre caracteres.

Dessa forma, para verificar a significância da correlação genética ($r_g$), sob a hipótese de que a verdadeira correlação (parâmetro) é nula, esses autores apresentam um teste t, dado por

$$t = r_g \Big/ \sqrt{\hat{var}(r_g)},$$

com aproximadamente n – 2 = $g_1$ – 1 graus de liberdade, sendo n o número de tratamentos genéticos avaliados e $g_1$ o número de graus de liberdade relativos a tratamentos.

A variância estimada $\hat{var}(r_g)$ é obtida sob a hipótese da nulidade, por meio da expressão

$$\hat{var}(r_g) = \frac{[1+(b-1)t_1][1+(b-2)t_2+r_f^2]}{g_1 b^2 t_1 t_2} + \frac{[(1-t_1)(1-t_2)+r_f^2]}{g_2 b^2 t_1 t_2},$$

em que

$g_1$, $g_2$: número de graus de liberdade relativos a tratamentos e resíduo, respectivamente;

*b*: número de repetições do ensaio;

$$t_1 = \frac{\hat{\sigma}_{gx}^2}{\hat{\sigma}_{gx}^2 + \hat{\sigma}_{ex}^2}; \qquad t_2 = \frac{\hat{\sigma}_{gy}^2}{\hat{\sigma}_{gy}^2 + \hat{\sigma}_{ey}^2} \quad \text{e}$$

$$r_f = \frac{\hat{\sigma}_{gxy} + \hat{\sigma}_{exy}}{[(\hat{\sigma}_{gx}^2 + \hat{\sigma}_{ex}^2)(\hat{\sigma}_{gy}^2 + \hat{\sigma}_{ey}^2)]^{0,5}},$$

onde

$\hat{\sigma}_{gxy}$: estimador da covariância genotípica entre os caracteres *X* e *Y*;

$\hat{\sigma}_{exy}$ : estimador da covariância ambiental entre os caracteres *X* e *Y*;

$\hat{\sigma}^2_{gx}$ e $\hat{\sigma}^2_{gy}$ : estimador das variâncias genotípicas dos caracteres *X* e *Y,* respectivamente;

$\hat{\sigma}^2_{ex}$ e $\hat{\sigma}^2_{ey}$ : estimador das variâncias ambientais dos caracteres *X* e *Y,* respectivamente.

## 2.2. Análise de trilha

Uma melhor compreensão das causas envolvidas nas associações entre caracteres é dada pela análise de trilha ou “path analysis”, desenvolvida pelo geneticista Sewall Wright em 1918-1921, segundo Johnson e Wichern (1992). Esse método consiste no desdobramento das correlações em efeitos diretos e indiretos, permitindo medir a influência direta de uma variável sobre a outra. De acordo com Cruz et al. (2004), as estimativas dos efeitos diretos e indiretos são obtidas por meio de equações de regressão, em que as variáveis são previamente padronizadas.

A análise de trilha, apesar de envolver princípios de regressão, é, em essência, um estudo da decomposição do coeficiente de correlação, permitindo avaliar se a relação entre duas variáveis é de causa e efeito ou se é determinada pela influência de outra(s) variável(is). A análise de trilha pode ser feita a partir de correlações fenotípicas, genotípicas ou ambientais (Cruz, 2006).

Conforme Johnson e Wichern (1992), a análise de trilha consiste em duas partes:

a) estabelecimento de um modelo de relacionamento de causa e efeito, ou diagrama causal, entre as variáveis;
b) decomposição das correlações observadas em um conjunto de termos denominado coeficientes de trilha, os quais representam os caminhamentos (trilhas) simples e complexos.

Assim, é possível medir os efeitos diretos e indiretos de uma variável sobre a outra.

### 2.2.1. Estimação dos coeficientes de trilha (efeitos diretos e indiretos)

Considerando-se, para efeito ilustrativo, uma variável básica *Y* e três variáveis explicativas ($X_1, X_2, X_3$), que se relacionam por meio do seguinte modelo, conforme descrito por Li (1975),

$$Y = b_{YX_1}(X_1) + b_{YX_2}(X_2) + b_{YX_3}(X_3) + \varepsilon .$$

Subtraindo-se a média de cada variável, tem-se

$$Y - \overline{Y} = b_{YX_1}(X_1 - \overline{X}_1) + b_{YX_2}(X_2 - \overline{X}_2) + b_{YX_3}(X_3 - \overline{X}_3) + \varepsilon .$$

Dividindo-se ambos os membros pelo desvio-padrão da variável básica ($\sigma_Y$), tem-se

$$\frac{Y - \overline{Y}}{\sigma_Y} = \frac{b_{YX_1}(X_1 - \overline{X}_1)}{\sigma_Y} + \frac{b_{YX_2}(X_2 - \overline{X}_2)}{\sigma_Y} + \frac{b_{YX_3}(X_3 - \overline{X}_3)}{\sigma_Y} + \frac{(\varepsilon - \overline{\varepsilon})}{\sigma_Y}.$$

Multiplicando-se e dividindo-se o 2º membro da expressão acima pelo desvio-padrão de cada variável explicativa ($\sigma_{X_1}, \sigma_{X_2}, \sigma_{X_2}$, respectivamente), da seguinte maneira

$$\frac{Y - \overline{Y}}{\sigma_Y} = \frac{b_{YX_1}(X_1 - \overline{X}_1)}{\sigma_Y}\frac{\sigma_{X_1}}{\sigma_{X_1}} + \frac{b_{YX_2}(X_2 - \overline{X}_2)}{\sigma_Y}\frac{\sigma_{X_2}}{\sigma_{X_2}} + \frac{b_{YX_3}(X_3 - \overline{X}_3)}{\sigma_Y}\frac{\sigma_{X_3}}{\sigma_{X_3}} + \frac{(\varepsilon - \overline{\varepsilon})}{\sigma_Y}\frac{\sigma_\varepsilon}{\sigma_\varepsilon}.$$

Dessa expressão, obtém-se

$$y = p_{yx_1}x_1 + p_{yx_2}x_2 + p_{yx_3}x_3 + p_\varepsilon u ,$$

em que

$y = \dfrac{Y - \overline{Y}}{\sigma_Y}$;

$$x_i = \frac{(X_i - \overline{X}_i)}{\sigma_{X_i}};$$

$$u = \frac{(\varepsilon - \overline{\varepsilon})}{\sigma_{\varepsilon}};$$

$$p_{\varepsilon=} \frac{\sigma_{\varepsilon}}{\sigma_Y};\ \text{e}$$

$$p_{yi=} \frac{b_{YX_i}\sigma_{X_i}}{\sigma_Y}.$$

Sendo

$$V(y) = V(p_{yx_1}x_1) + V(p_{yx_2}x_2) + V(p_{yx_3}x_3) + V(p_{\varepsilon}u) + 2Cov(p_{yx_1}x_1, p_{yx_2}x_2) + 2Cov(p_{yx_1}x_1, p_{yx_3}x_3) + 2Cov(p_{yx_1}x_1, p_{\varepsilon}u) + 2Cov(p_{yx_2}x_2, p_{yx_3}x_3) + 2Cov(p_{yx_2}x_2, p_{\varepsilon}u) + 2Cov(p_{yx_3}x_3, p_{\varepsilon}u).$$

Ou de maneira mais simplificada

$$V(y) = p_{yx_1}^2 V(x_1) + p_{yx_2}^2 V(x_2) + p_{yx_3}^2 V(x_3) + p_{\varepsilon}^2 V(u) + 2p_{yx_1}p_{yx_2}Cov(x_1, x_2) + 2p_{yx_1}p_{yx_3}Cov(x_1, x_3) + 2p_{yx_1}p_{\varepsilon}Cov(x_1, u) + 2p_{yx_2}p_{yx_3}Cov(x_2, x_3) + 2p_{yx_2}p_{\varepsilon}Cov(x_2, u) + 2p_{yx_3}p_{\varepsilon}Cov(x_3, u).$$

Sendo

$$i)\ x_i = \frac{(X_i - \overline{X}_i)}{\sigma_{X_i}},$$

tem-se que $V(x_i) = V\left(\frac{X_i - \overline{X}}{\sigma_{X_i}}\right)$, então, $V(x_i) = \frac{V(X_i)}{\sigma_{X_i}^2} = 1.$

De modo análogo, tem-se $V(y) = 1$ e $V(u) = 1$.

$$ii)\ Cov(y,x_i) = Cov\left(\frac{Y-\overline{Y}}{\sigma_Y}, \frac{X_i-\overline{X}}{\sigma_{X_i}}\right)$$

$$= \frac{1}{\sigma_Y \sigma_{X_i}} Cov(Y,X_i) - Cov(Y,\overline{X}) - Cov(\overline{Y},X_i) + Cov(\overline{Y},\overline{X})$$

$$= \frac{1}{\sigma_Y \sigma_{X_i}} Cov(Y,X_i) = r_{YX_i}.$$

$$iii)\ Cov(x_i,x_j) = Cov\left(\frac{X_i-\overline{X}}{\sigma_{X_i}}, \frac{X_j-\overline{X}}{\sigma_{X_j}}\right)$$

$$= \frac{1}{\sigma_{X_i} \sigma_{X_j}} Cov(X_i,X_j) - Cov(X_i,\overline{X}) - Cov(\overline{X},X_j) + Cov(\overline{X},\overline{X})$$

$$= \frac{1}{\sigma_{X_i} \sigma_{X_j}} Cov(X_i,X_j) = r_{X_iX_j}.$$

$$iv)\ Cov(u,x_i) = Cov\left(\frac{\varepsilon-\overline{\varepsilon}}{\sigma_\varepsilon}, \frac{X_i-\overline{X}}{\sigma_{x_i}}\right)$$

$$= \frac{1}{\sigma_\varepsilon \sigma_{X_i}} Cov(\varepsilon,X_i) - Cov(\varepsilon,\overline{X}) - Cov(\overline{\varepsilon},X_i) + Cov(\overline{\varepsilon},\overline{X})$$

$$= 0.$$

É possível verificar as seguintes relações:

$$V(y) = p_{yx_1}^2 + p_{yx_2}^2 + p_{yx_3}^2 + 2p_{yx_1}p_{yx_2}r_{12} + 2p_{yx_1}p_{yx_3}r_{13} + 2p_{yx_2}p_{yx_3}r_{23} + p_\varepsilon^2;$$

$$V(y) = V(\hat{y}) + p_\varepsilon^2;$$

$$V(\hat{y}) = p_{yx_1}^2 + p_{yx_2}^2 + p_{yx_3}^2 + 2p_{yx_1}p_{yx_2}r_{12} + 2p_{yx_1}p_{yx_3}r_{13} + 2p_{yx_2}p_{yx_3}r_{23} \quad \text{e}$$

$$Cov(y,x_1) = r_{YX_1} = Cov(p_{yx_1}x_1 + p_{yx_2}x_2 + p_{yx_3}x_3 + p_\varepsilon u, x_1)$$

$$r_{YX_1} = p_{yx_1}Cov(x_1,x_1) + p_{yx_2}Cov(x_1,x_2) + p_{yx_3}Cov(x_1,x_3) + p_\varepsilon Cov(u,x_1)$$

$$r_{YX_1} = p_{yx_1} + p_{yx_2}r_{12} + p_{yx_3}r_{13}.$$

De maneira análoga são determinadas as demais correlações:

$$r_{YX_2} = p_{yx_1}r_{12} + p_{yx_2} + p_{yx_3}r_{23} \quad \text{e}$$

$$r_{YX_3} = p_{yx_1} r_{13} + p_{yx_2} r_{23} + p_{yx_3}.$$

Da expressão

$$V(\hat{y}) = p_{yx_1}^2 + p_{yx_2}^2 + p_{yx_3}^2 + 2p_{yx_1}p_{yx_2}r_{12} + 2p_{yx_1}p_{yx_3}r_{13} + 2p_{yx_2}p_{yx_3}r_{23}$$

pode ser estimado o coeficiente de determinação do modelo causal ($R^2$), que mede os efeitos das variáveis explicativas sobre a variável principal.

O coeficiente de determinação ($R^2$) é dado por:

$$R^2 = \frac{\text{SQRegressão}}{\text{SQTotal}}$$

Como $\text{SQRegressão} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = V(\hat{y})$ e

$\text{SQTotal} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = V(y)$, o coeficiente de determinação pode ser expresso por:

$$R^2 = \frac{\text{V}(\hat{\text{y}})}{\text{V(y)}}.$$

Dado que $V(y) = 1$, então,

$$R^2 = p_{yx_1}^2 + p_{yx_2}^2 + p_{yx_3}^2 + 2p_{yx_1}p_{yx_2}r_{12} + 2p_{yx_1}p_{yx_3}r_{13} + 2p_{yx_2}p_{yx_3}r_{23}.$$

Pode ser estimado também o efeito da variável residual sobre a variável principal, fazendo-se:

$$V(y) = p_{yx_1}^2 + p_{yx_2}^2 + p_{yx_3}^2 + 2p_{yx_1}p_{yx_2}r_{12} + 2p_{yx_1}p_{yx_3}r_{13} + 2p_{yx_2}p_{yx_3}r_{23} + p_{\varepsilon}^2$$

$$V(y) = R^2 + p_{\varepsilon}^2$$

$$1 = R^2 + p_{\varepsilon}^2$$

$$p_{\varepsilon}^2 = 1 - R^2$$

$$p_{\varepsilon} = \sqrt{1 - R^2}.$$

Os estimadores dos coeficientes de trilha, obtidos pelo processo descrito acima, são de mínimos quadrados, pois provêm das equações normais $X'X\hat{\beta} = X'Y$, cujas matrizes são derivadas do modelo $y = p_{yx_1}x_1 + p_{yx_2}x_2 + p_{yx_3}x_3 + p_{\varepsilon}u$ (Cruz et al., 2004). No sistema de equações normais, $X'X$ é a matriz de correlações genotípicas ou fenotípicas entre as variáveis explicativas do modelo; $\hat{\beta}$ é o vetor dos estimadores dos coeficientes de trilha; $X'Y$ é a matriz de correlações genotípicas ou fenotípicas entre a variável principal e cada variável explicativa do modelo, como segue:

$$X'Y = \begin{bmatrix} r_{YX_1} \\ r_{YX_2} \\ r_{YX_3} \end{bmatrix},\ X'X = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} \text{ e } \hat{\beta} = \begin{bmatrix} p_{yx_1} \\ p_{yx_2} \\ p_{yx_3} \end{bmatrix}$$

$$\begin{bmatrix} r_{YX_1} \\ r_{YX_2} \\ r_{YX_3} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} \begin{bmatrix} p_{yx_1} \\ p_{yx_2} \\ p_{yx_3} \end{bmatrix}$$

Os resultados da análise de trilha permitem melhor interpretação por meio do diagrama apresentado na Figura 1.

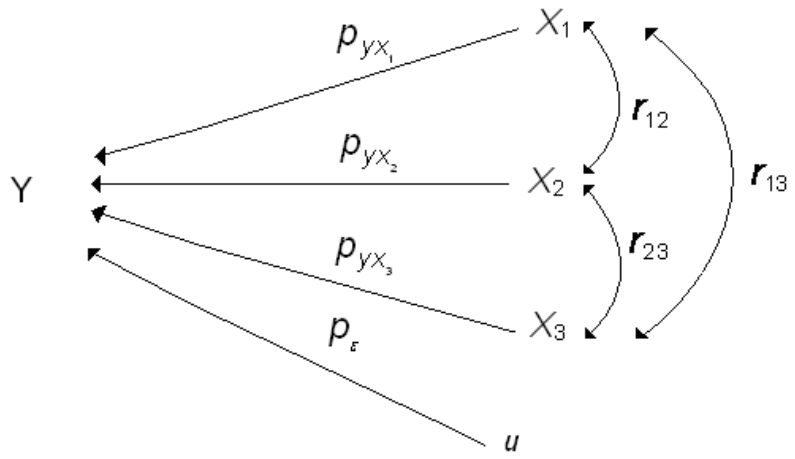


**Figura 1.** Diagrama causal ilustrativo dos efeitos das variáveis explicativas ( $X_1, X_2, X_3$ ) e residual (*u*) sobre a variável dependente *Y*.

## 2.3. Multicolinearidade

### 2.3.1. Conceito e causas

O termo multicolinearidade, segundo Gujarati (2000), foi criado por Ragnar Frisch em 1934. Significava originalmente a existência de uma "perfeita" (ou exata) relação linear entre algumas ou todas as variáveis explicativas de um modelo de regressão.

Segundo Neter et al. (1990) e Kutner et al. (2005), a multicolinearidade ocorre quando existe algum nível de inter-relação entre as variáveis independentes do modelo de regressão linear múltipla. Algumas vezes o termo multicolinearidade é utilizado apenas nos casos em que a correlação entre as variáveis é muito alta. De acordo com Belsley et al. (1980), os termos colinearidade, multicolinearidade, e mal-condicionamento são todos usados para denotar a mesma situação.

Algumas questões frequentemente relacionadas à regressão múltipla incluem o estabelecimento da importância relativa e da magnitude do efeito das variáveis explicativas sobre a variável dependente; a identificação de

preditores que poderiam ser eliminados do modelo devido ao pequeno ou nenhum efeito sobre a variável resposta; a possibilidade de inclusão de variáveis ainda não participantes do modelo (Neter et al, 1990; Kutner et al., 2005).

Segundo esses autores, na ausência de multicolinearidade, respostas relativamente simples podem ser dadas a essas questões. No entanto, em muitas situações não experimentais, tais como, em finanças, economia, ciências sociais e biológicas, as variáveis independentes tendem a ser correlacionadas entre si e com outras variáveis que estão relacionadas com a variável dependente, mas que não estão incluídas no modelo.

Quanto às possíveis fontes de multicolinearidade, Montgomery e Peck (1992) destacam as seguintes:

1. O método utilizado na obtenção dos dados;
2. Restrições sobre o modelo ou sobre a população;
3. Especificação do modelo;
4. Modelos super parametrizados.

De acordo com Belsley et al. (1980), intuitivamente, pode-se entender o virtual dano resultante de dados provenientes de variáveis multicolineares, ao constatar que tais variáveis não fornecem informações muito diferentes daquelas já inerentes em outras, o que torna difícil inferir sobre as influências individuais dessas variáveis independentes sobre a resposta.

Em muitos casos, quando é envolvido um grande número de variáveis ou não há conhecimento prévio da associação entre elas, resultados inapropriados, gerados pela multicolinearidade, podem ser interpretados, levando a conclusões que não seriam as mais pertinentes (Cruz e Carneiro, 2006).

Como observado por Kutner et al. (2005), a multicolinearidade tende a produzir estimativas de mínimos quadrados $\hat{\beta}_j$ com variâncias muito grandes. Considerando-se o modelo de regressão linear múltipla com as variáveis padronizadas $x_1, x_2$ e $y$ :

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon ,$$

onde as equações normais de mínimos quadrados são

$$(X'X)\hat{\beta} = X'y$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix},$$

e onde $r_{12}$ é a correlação entre $x_1$ e $x_2$ e $r_{jy}$ é a correlação entre $x_j$ e $y$, $j = 1, 2$. A inversa de $(X'X)$ é

$$C = (X'X)^{-1} = \begin{bmatrix} \frac{1}{(1-r_{12}^2)} & \frac{-r_{12}}{(1-r_{12}^2)} \\ \frac{-r_{12}}{(1-r_{12}^2)} & \frac{1}{(1-r_{12}^2)} \end{bmatrix}$$

e as estimativas dos coeficientes de regressão são

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{(1-r_{12}^2)}, \qquad \hat{\beta}_2 = \frac{r_{2y} - r_{12}r_{1y}}{(1-r_{12}^2)}.$$

Se existe multicolinearidade forte entre $x_1$ e $x_2$, então o coeficiente de correlação $r_{12}$ será próximo de mais ou menos 1. Assim, de $(X'X)^{-1}$ tem-se que

$$|r_{12}| \to 1, \quad V(\hat{\beta}_j) = C_{jj}\sigma^2 \to \infty$$

e

$$COV(\hat{\beta}_1, \hat{\beta}_2) = C_{12}\sigma^2 \to \pm\infty \text{ , se } r_{12} \to +1 \text{ ou } r_{12} \to -1.$$

Logo, a multicolinearidade forte entre $x_1$ e $x_2$ resulta em variâncias e covariâncias grandes para os estimadores de mínimos quadrados dos coeficientes de regressão.

Para o caso de mais de duas variáveis, a multicolinearidade produz efeitos similares. Os elementos da diagonal principal da matriz $C = (X'X)^{-1}$ são

$$C_{jj} = \frac{1}{1-R_j^2}, \qquad j = 1, 2, ...., p,$$

onde o $R_j^2$ é o coeficiente de determinação múltipla da regressão de $x_j$ sobre as variáveis independentes restantes. Assim, uma forte

multicolinearidade entre $x_j$ e qualquer subconjunto das demais variáveis independentes, implicará em $R_j^2$ próximo de 1. Como a variância de $\hat{\beta}_j$ é

$$V(\hat{\beta}_j) = C_{jj}\sigma^2 = (1-R_j^2)^{-1}\sigma^2,$$

a multicolinearidade pode produzir uma variância da estimativa de mínimos quadrados do coeficiente de regressão $\hat{\beta}_j$ muito grande. Em geral, a covariância de $\hat{\beta}_i$ e $\hat{\beta}_j$ deverá também ser grande, se as variáveis $x_i$ e $x_j$ estão envolvidas em uma relação de multicolinearidade.

A multicolinearidade também tende a produzir estimativas de mínimos quadrados dos coeficientes de regressão muito grandes, afetando o quadrado da distância entre o estimador de mínimos quadrados $\hat{\beta}$ e o verdadeiro vetor de parâmetros $\beta$ (Hoerl e Kennard, 1970a), assim,

$$L = \hat{\beta} - \beta$$
$$L^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta).$$

O valor esperado do quadrado da distância, $E(L_1^2)$, é

$$\begin{aligned} E(L^2) &= E(\hat{\beta} - \beta)'(\hat{\beta} - \beta) \\ &= \sum_{j=1}^{p} E(\hat{\beta}_j - \beta_j)^2 \\ &= \sum_{j=1}^{p} V(\hat{\beta}_j) \\ &= \sigma^2 Tr(X'X)^{-1}, \end{aligned}$$

onde o traço $(Tr)$ da matriz é apenas a soma dos elementos da diagonal principal de $(X'X)$. Dessa forma,

$$E(L^2) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j},$$

onde $\lambda_j > 0,\ j = 1, 2, ..., p$ são os autovalores da matriz $(X'X)$, considerada na forma de correlações. Assim, se a matriz $(X'X)$ é mal condicionada em razão da multicolinearidade, pelo menos um dos $\lambda_j$ será muito pequeno,

implicando que a distância da estimativa de mínimos quadrados $\hat{\beta}$ ao parâmetro $\beta$ poderá ser muito grande. Equivalentemente, tem-se que

$$E(L^2) = E(\hat{\beta} - \beta)'(\hat{\beta} - \beta)$$
$$= E(\hat{\beta}'\hat{\beta} - 2\hat{\beta}'\beta + \beta'\beta)$$

ou

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 Tr(X'X)^{-1}.$$

Isto é, o vetor $\hat{\beta}$ é geralmente maior que o vetor $\beta$. Dessa forma, o método de mínimos quadrados produz estimativas dos coeficientes de regressão que podem ser muito grandes em valor absoluto.

Mandel (1982) argumenta que apesar do método dos mínimos quadrados, em geral, produzir estimativas dos parâmetros individuais do modelo demasiadamente instáveis na presença de multicolinearidade, não implica, necessariamente, que o modelo não seja bom preditor. Se as predições estiverem restritas a regiões do espaço das variáveis independentes onde a multicolinearidade se verifica, o modelo ajustado frequentemente produzirá predições satisfatórias.

### 2.3.2. Efeitos da multicolinearidade sobre a análise de trilha

Sendo a análise de trilha uma forma de estudo de regressão, com base em matrizes de correlações, pode-se deduzir, por analogia, o virtual dano que a multicolinearidade causa ao processo de estimação dos coeficientes (Carvalho, 1995).

Assim, existindo multicolinearidade em níveis considerados moderados a severos entre um conjunto de variáveis explicativas, as variâncias associadas a certos estimadores, como, por exemplo, os coeficientes de trilha que medem os efeitos diretos de variáveis explicativas sobre uma variável principal, podem atingir valores excessivamente elevados, sendo evidência de serem as estimativas pouco confiáveis ou sem nenhuma coerência com o fenômeno biológico em estudo (Cruz e Carneiro, 2006). Coimbra et al. (2005),

por exemplo, ao estudarem o grau de multicolinearidade sobre a análise de trilha em canola, concluíram que a aplicação da análise de trilha sobre o grau de multicolinearidade severa produz resultados sem nenhuma importância biológica para o melhorista de plantas.

### 2.3.3. Diagnóstico de multicolinearidade

O diagnóstico de multicolinearidade dito eficiente é aquele que, além de informar sobre a existência da multicolinearidade, apresenta o grau de severidade e identifica as variáveis envolvidas no problema (Montgomery e Peck, 1992).

Dentre os procedimentos mais comumente utilizados para detectar a multicolinearidade, conforme indicam vários autores, estão os seguintes:

#### 2.3.3.1. Análise da matriz de correlação

Para Belsley et al. (1980), uma medida muito simples para examinar a multicolinearidade consiste na análise dos elementos não-diagonais ($r_{ij}$) na matriz ($X'X$) de trabalho.

Se as variáveis independentes ($x_i, x_j$) apresentarem dependência linear aproximada entre si, então $r_{ij}$, tomado em valor modular $|r_{ij}|$ será próximo à unidade. De acordo com esses autores, enquanto um alto coeficiente de correlação entre duas variáveis independentes pode de fato apontar para um possível problema de colinearidade, a ausência de correlações altas não pode ser vista como evidência de nenhum problema. É possível três ou mais variáveis independentes estarem envolvidas em uma relação de multicolinearidade, sem que quaisquer pares dessas variáveis sejam altamente correlacionadas. Tal situação não é diagnosticada pelo exame da matriz de correlação.

### 2.3.3.2. Fatores de inflação da variância

Os elementos diagonais $C_{jj} = \frac{1}{1-R_j^2}$, $j = 1, 2, ...., p$ da matriz $C = (X'X)^{-1}$, na forma de correlação, denominados por Marquardt, em 1970, de fatores de inflação da variância (VIFs) constituem um importante diagnóstico da multicolinearidade, conforme Belsley et al. (1980). Esses mesmos autores observam que, como a variância do j-ésimo coeficiente de regressão de mínimos quadrados ($\hat{\beta}_j$) é $V(\hat{\beta}_j) = C_{jj}\sigma^2 = (1-R_j^2)^{-1}\sigma^2$, é possível considerar $C_{jj}$ como o fator pelo qual a variância de $\hat{\beta}_j$ é aumentada devido à dependência linear entre as variáveis.

De acordo com Neter et al. (1990), o maior valor de VIF entre todas as variáveis é frequentemente usado como um indicador da severidade da multicolinearidade, e a ocorrência de qualquer VIF com valor superior a 10 constitui indicativo de que a multicolinearidade pode estar influenciando indevidamente as estimativas de mínimos quadrados.

### 2.3.3.3. Análise dos autovalores e autovetores da matriz

Segundo Belsley et al. (1980), Kendall (1957) e Silvey (1969) as raízes características ou autovalores de $X'X$, denotados por $\lambda_1, \lambda_2, ..., \lambda_p$, podem ser usados para medir o grau de multicolinearidade nos dados. Se existe uma ou mais dependências lineares aproximadas nesses dados, um ou mais autovalores serão pequenos.

Montgomery e Peck (1992) notam que alguns autores preferem examinar o número de condição (*NC*) de $X'X$, definido como a razão entre o maior e o menor autovalor dessa matriz, ou seja,

$$NC = \frac{\lambda_{\max}}{\lambda_{mín}}$$

Como critério para a classificação da multicolinearidade de uma matriz, esses autores argumentam que, em geral, se o número de condição é menor que 100, não existe problema sério com a multicolinearidade; para um valor de *NC* entre 100 e 1000 implica em multicolinearidade moderada a forte; se o número de condição for maior que 1000, constitui indício de multicolinearidade severa.

A análise dos autovalores também pode ser usada para identificar a natureza da dependência linear aproximada nos dados, de acordo, ainda, com Montgomery e Peck (1992). A matriz $X'X$ pode ser decomposta como

$$X'X = T\Lambda T',$$

sendo

$\Lambda$: matriz diagonal *pxp,* cujos elementos da diagonal são os autovalores $\lambda_j$ $(j = 1,2,...,p)$ de $X'X$, ou seja,

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & 0 & \lambda_p \end{bmatrix};$$

*T*: matriz ortogonal *pxp*, cujas colunas $(t_1, t_2, ..., t_p)$ são os autovetores normalizados de $X'X$, isto é,

$$T = [t_1 \quad t_2 \quad ... \quad t_p] = \begin{bmatrix} t_{11} & t_{12} & ... & t_{1p} \\ t_{21} & t_{22} & ... & t_{2p} \\ ... & ... & ... & ... \\ t_{p1} & t_{p2} & ... & t_{pp} \end{bmatrix}.$$

Assim, se $X'X = T\Lambda T'$ e sabendo $T'T = I$, tem-se:

$$T'(X'X)T = \Lambda,$$

ou ainda,

$$t_j'(X'X)t_j = \lambda_j,$$

e

$$t_j'(X'X)t_k = 0 \quad (k \neq j).$$

Se um autovalor $(\lambda_j)$ é próximo de zero, indicando uma dependência linear entre as observações, os elementos do autovetor associado a esse autovalor descrevem a natureza dessa dependência linear.

### 2.3.3.4. Decomposição em valores singulares

De acordo com Lawson e Hanson (1974) e Belsley et al. (1980), qualquer matriz *X nxp*, sendo *n* observações e *p* variáveis, pode ser decomposta da seguinte forma:

$$X = UDT',$$

onde *U* é uma matriz *nxp* cujas colunas são os autovetores associados aos *p* autovalores não nulos de $X'X$, *T* é uma matriz *pxp* dos autovetores normalizados de $X'X$, $U'U = T'T = I$ e *D* é uma matriz diagonal *pxp* com elementos diagonais não-negativos $\mu_j$, $j = 1,2,...,p$, denominados valores singulares de *X*. Então, $X = UDT'$ é a forma de decomposição de *X* em seus valores singulares.

Para Belsley et al. (1980) a decomposição em valores singulares está estritamente relacionada com os conceitos de autovalores e autovetores, dado que

$$X'X = (UDT')'UDT' = TD^2T' = T\Lambda T',$$

de modo que o quadrado dos valores singulares de *X* são os autovalores de $X'X$.

Belsley et al. (1980) notam que a matriz *X* mal-condicionada afeta o tamanho dos valores singulares, havendo um valor singular pequeno para cada dependência linear aproximada. A magnitude do mal-condicionamento depende de quão pequeno é o valor singular mínimo em relação ao valor singular máximo $\mu_{máx}$. Dessa forma, Belsley et al. (1980) definem o índice de condição da matriz *X* como:

$$\eta_k = \frac{\mu_{máx}}{\mu_k}, \quad j = 1,2,...p.$$

Assim, o maior valor para $\eta_k$ representa o número de condição da matriz *X*. Segundo Montgomery e Peck (1992), a vantagem adicional dessa técnica é que os algoritmos para gerar a decomposição dos valores singulares são numericamente mais estáveis em relação àqueles da análise dos autovalores e autovetores, embora, na prática, não seja, provavelmente, uma desvantagem severa a preferência pela análise dos autovalores e autovetores.

Sabendo que $\sigma^2(X'X)^{-1}$ é a matriz de variâncias e covariâncias do estimador de mínimos quadrados $\hat{\beta} = (X'X)^{-1}X'y$, onde $\sigma^2$ é a variância comum dos componentes de $\varepsilon$ no modelo linear $y = X\beta + \varepsilon$, e utilizando a decomposição em valores singulares, $X = UDT'$, a matriz de variância e covariância de $\hat{\beta}$, isto é, $V(\hat{\beta})$, pode ser escrita, conforme Belsley et al. (1980), como:

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1} = \sigma^2 T\Lambda^{-1}T'$$

e a variância do j-ésimo coeficiente de regressão é o j-ésimo elemento diagonal dessa matriz, ou

$$V(\hat{\beta}_k) = \sigma^2 \sum_{i=1}^{p} \frac{t_{kj}^2}{\mu_j^2} = \sigma^2 \sum_{i=1}^{p} \frac{t_{kj}^2}{\lambda_j}.$$

À exceção de $\sigma^2$, o j-ésimo elemento diagonal de $T\Lambda^{-1}T'$ é o j-ésimo fator de inflação da variância, ou seja

$$VIF_j = \sum_{i=1}^{p} \frac{t_{kj}^2}{\mu_j^2} = \sum_{i=1}^{p} \frac{t_{kj}^2}{\lambda_j}.$$

Claramente, a ocorrência de um ou mais valores singulares pequenos, ou autovalores pequenos, pode inflacionar dramaticamente a variância de $\hat{\beta}_j$.

**2.3.3.5. Determinante de $X'X$**

Conforme Montgomery e Peck (1992), o determinante de $X'X$ pode ser usado como um índice de multicolinearidade. Desde que a matriz $X'X$ esteja na forma de correlação, a possível variação dos valores do determinante é $0 \leq |X'X| \leq 1$. Se $|X'X| = 1$, as variáveis independentes são ortogonais. Se $|X'X| = 0$, existe uma dependência linear completa entre essas variáveis. À medida que $|X'X|$ se aproxima de zero, a multicolinearidade se torna mais intensa. Esses autores ainda observam que, embora esta técnica seja facilmente aplicada, ela não fornece qualquer informação sobre a origem da multicolinearidade.

É importante notar que esse intervalo do determinante não se aplica ao caso do uso da matriz de correlação genotípica (Peternelli, 2009).

### 2.3.4. Processos alternativos ao de mínimos quadrados quando ocorre a multicolinearidade

Para solucionar os problemas adversos da multicolinearidade, várias técnicas têm sido propostas, tais como, a obtenção de dados adicionais, a reespecificação do modelo e o uso de outros métodos de estimação de mínimos quadrados que são especificamente planejados para combater os problemas induzidos pela multicolinearidade (Montgomery e Peck, 1992).

Esses autores observam, no entanto, que o uso desses procedimentos nem sempre é possível ou mesmo viável. Quando a multicolinearidade é devido a restrições sobre o modelo ou sobre a população, a coleta de dados adicionais é uma solução pouco recomendável. Em relação à eliminação de variáveis, apesar de ser uma técnica, em geral, altamente efetiva, poderá não estabelecer uma solução satisfatória se as variáveis retiradas do modelo tiverem um grande poder explicativo em relação à resposta, prejudicando sobremaneira o poder de predição do modelo. E ainda, muitos dos procedimentos de seleção de variáveis são seriamente distorcidos pela multicolinearidade, podendo resultar em um modelo final com grau de multicolinearidade não muito inferior ao modelo original.

Os métodos alternativos ao de mínimos quadrados, especificamente planejados para combater os problemas induzidos pela multicolinearidade, fornecem estimadores tendenciosos, mas, conforme Gunst e Mason (1977) apresentam, em geral, melhor desempenho quando comparados aos estimadores de mínimos quadrados. Dentre esses procedimentos, são citados dois métodos: um denominado “ridge regression” (regressão em crista ou em cumeeira, conforme Cruz e Carneiro, 2006), originalmente proposto por Hoerl e Kennard (1970a,b); e o outro denominado regressão em componentes principais, segundo Chatterjee e Price (1977).

### 2.3.4.1. Regressão em crista

Esse procedimento é denominado regressão em crista devido à similaridade ao método de análise em crista originalmente desenvolvido por Hoerl em 1959 para descrever o comportamento de superfícies de resposta multidimensionais (Hoerl, 1985).

De acordo com Cruz e Carneiro (2006), o método de regressão em crista se baseia em obter estimativas de coeficientes de regressão a partir de uma versão ligeiramente modificada das equações normais. Especificamente, o estimador em cristas $\hat{\beta}^*$ é definido como a solução para

$$(X'X + Ik)\hat{\beta}^* = X'y$$

ou

$$\hat{\beta}^* = (X'X + Ik)^{-1}X'y\,,$$

onde $0 \leq k \leq 1$, uma vez que $X'X$ se encontra na forma de correlações.

Hoerl e Kennard (1970a) destacam as seguintes propriedades aplicáveis à regressão em crista:

1. O estimador em crista $\hat{\beta}^*$ é uma transformação linear do estimador de mínimos quadrados $\hat{\beta}$, assim,

$$\hat{\beta}^* = (X'X + Ik)^{-1}X'y$$
$$E(\hat{\beta}^*) = (X'X + Ik)^{-1}X'E(y)$$
$$E(\hat{\beta}^*) = (X'X + Ik)^{-1}X'E(X\beta + \varepsilon)$$
$$E(\hat{\beta}^*) = (X'X + Ik)^{-1}X'X\beta + E(\varepsilon)$$
$$E(\hat{\beta}^*) = (X'X + Ik)^{-1}X'X\beta + \phi$$
$$E(\hat{\beta}^*) = (X'X + Ik)^{-1}X'X\beta \neq \beta.$$

Logo, $\hat{\beta}^*$ é um estimador tendencioso de $\beta$, visto que

$$\hat{\beta}^* = \omega\hat{\beta},$$

em que

$$\omega = (X'X + Ik)^{-1}X'X$$
$$\omega = (X'X + Ik)^{-1}(X'X + Ik - Ik) = (X'X + Ik)^{-1}(X'X + Ik) - Ik(X'X + Ik)^{-1}$$
$$\omega = I - k(X'X + Ik)^{-1}.$$
$$E(\hat{\beta}^*) = [I - k(X'X + Ik)^{-1}]\beta = [\beta - k(X'X + Ik)^{-1}\beta.$$

Logo,

$$\text{Viés}(\beta^*) = E(\hat{\beta}^*) - \beta = -k(X'X + Ik)^{-1}\beta.$$

Alguns fatos:

i) $E[\hat{\beta}^* - E(\hat{\beta}^*)]$

Como

$\hat{\beta}^* = (X'X + Ik)^{-1}X'y$ e

$E(\hat{\beta}^*) = (X'X + Ik)^{-1}X'X\beta,$

assim,

$$\hat{\beta}^* - E(\hat{\beta}^*) = (X'X + Ik)^{-1}X'(X\beta + \varepsilon) - (X'X + Ik)^{-1}X'X\beta$$
$$\hat{\beta}^* - E(\hat{\beta}^*) = (X'X + Ik)^{-1}X'\varepsilon.$$

Logo,

$$E[\hat{\beta}^* - E(\hat{\beta}^*)] = E[(X'X + Ik)^{-1}X'\varepsilon] = \phi$$

ii) $E\{[\beta - E(\hat{\beta}^*)]'[\beta - E(\hat{\beta}^*)]\}$

Como

$$\beta - E(\hat{\beta}^*) = \beta - \beta - k(X'X + Ik)^{-1}\beta$$

$$\beta - E(\hat{\beta}^*) = -k(X'X + Ik)^{-1}\beta,$$

assim

$$E\{[\beta - E(\hat{\beta}^*)]'[\beta - E(\hat{\beta}^*)]\} = k^2\beta'(X'X + Ik)^{-2}\beta = \sum_{j=1}^{p}[Vi\acute{e}s(\beta_j^*)]^2.$$

2. Covariância de $\hat{\beta}^*$

$$COV(\hat{\beta}^*) = E\{[\hat{\beta}^* - E(\hat{\beta}^*)][\hat{\beta}^* - E(\hat{\beta}^*)]'\}$$

$$COV(\hat{\beta}^*) = E\{[(X'X + Ik)^{-1}X'\varepsilon][(X'X + Ik)^{-1}X'\varepsilon]'\}$$

$$COV(\hat{\beta}^*) = (X'X + Ik)^{-1}X'E(\varepsilon\varepsilon')X(X'X + Ik)^{-1}$$

$$COV(\hat{\beta}^*) = (X'X + Ik)^{-1}X'X(X'X + Ik)^{-1}\sigma^2.$$

3. O erro quadrático médio de $\hat{\beta}^*$ é designado por $EQM(\hat{\beta}^*)$, cuja esperança é dada por

$$E[EQM(\hat{\beta}^*)] = E[(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)]$$

ou

$$\begin{aligned} E[EQM(\hat{\beta}^*)] &= E\{[(\hat{\beta}^* - E(\hat{\beta}^*)) - (\beta - E(\hat{\beta}^*))]'[\hat{\beta}^* - E(\hat{\beta}^*)) - \\ &\quad (\beta - E(\hat{\beta}^*))]\} \\ &= E[(\hat{\beta}^* - E(\hat{\beta}^*))'(\hat{\beta}^* - E(\hat{\beta}^*))] - 2E[(\hat{\beta}^* - E(\hat{\beta}^*))'(\beta - E(\hat{\beta}^*))] + \\ &\quad E[(\beta - E(\hat{\beta}^*))'(\beta - E(\hat{\beta}^*))]. \end{aligned}$$

Logo,

$$E[EQM(\hat{\beta}^*)] = \sum_{j=1}^{p}V(\beta_j^*) + \sum_{j=1}^{p}[\text{Viés } \beta^*]^2$$

$$= \sigma^2 Tr[(X'X + kI)^{-1}X'X(X'X + kI)^{-1}] + k^2\beta'(X'X + kI)^{-2}\beta$$

$$= \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j + k)^2} + k^2\beta'(X'X + kI)^{-2}\beta,$$

sendo

$$\sum_{j=1}^{p} V(\beta_j^*) = \sigma^2 Tr[(X'X + Ik)^{-1} X'X (X'X + Ik)^{-1} = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2},$$

onde $\lambda_1, \lambda_2, ..., \lambda_p$ são os autovalores de *X'X*.

Esta propriedade tem dois corolários importantes:

i) A variância de $\hat{\beta}^*$ é uma função decrescente de *k*;

ii) O viés é uma função crescente de *k*.

4. Se $\beta'\beta$ é limitado, então existe um valor de $k > 0$, tal que

$$EQM(\hat{\beta}^*) < EQM(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1},$$

propriedade esta denominada teorema da existência.

Segundo Montgomery e Peck (1992), a escolha correta da constante *k* tem sido objeto de muitas discussões acerca do emprego da regressão em crista, com diferentes procedimentos propostos por vários autores.

Hoerl e Kennard (1970a) sugerem que um valor apropriado de *k* pode ser determinado pela inspeção do traço de crista. O traço de crista é um diagrama dos elementos de $\hat{\beta}^*$ por *k*, para valores de *k* normalmente no intervalo [0,1].

De acordo com Montgomery e Peck (1992), se a multicolinearidade é severa, será evidente a instabilidade nos coeficientes de regressão pelo traço de crista. À medida que o valor de *k* aumenta, algumas estimativas em crista irão variar dramaticamente, e para algum valor de *k*, a estimativa em crista $\hat{\beta}^*$ será estável. Então, o objetivo é selecionar um valor razoavelmente pequeno de *k*, cujas estimativas em crista $\hat{\beta}^*$ são estáveis, o que certamente produzirá um conjunto de estimativas com um $EQM(\hat{\beta}^*)$ menor que as estimativas de mínimos quadrados.

Segundo Hoerl e Kennard (1970b), um valor adequado para *k* pode ser obtido pela observação do traço de crista e pela escolha de *k*, onde:

- Os coeficientes já se estabilizaram e terão o comportamento semelhante ao de um sistema ortogonal;

- Os coeficientes com sinais aparentemente incorretos em $k = 0$, serão mudados para o sinal apropriado;

- Os resíduos não serão muito elevados em relação ao resíduo obtido em $k = 0$.

Newell e Lee (1981) verificaram que a regressão em crista tem se mostrado uma alternativa conveniente, não somente superando o problema da inflação da variância e da instabilidade das estimativas dos coeficientes de regressão, como também, possibilitando ao pesquisador maior estudo das inter-relações entre as variáveis independentes envolvidas no modelo.

### 2.3.4.2. Regressão em componentes principais

A aplicação da análise de componentes principais é recomendada por Chatterjee e Price (1977), dentre outros, como procedimento alternativo ao de mínimos quadrados, para contornar os problemas de multicolinearidade apresentados pelos dados.

Os componentes principais podem ser obtidos, segundo Hocking (1976) e Montgomery e Peck (1992), considerando o modelo na forma canônica

$$y = Z\alpha + \varepsilon,$$

onde $Z = XT$, $\alpha = T'\beta$, $T'X'XT = Z'Z = \Lambda$ e $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_p)$ é uma matriz diagonal *pxp* dos autovalores de $X'X$ e *T* é uma matriz ortogonal *pxp,* cujas colunas são os autovetores associados a $\lambda_1, \lambda_2, ..., \lambda_p$. As colunas da matriz $Z = [Z_1, Z_2, ..., Z_p]$, que definem um novo conjunto de variáveis ortogonais, são denominadas componentes principais.

O estimador de mínimos quadrados de $\alpha$ é

$$\hat{\alpha} = (Z'Z)^{-1}Z'y = \Lambda^{-1}Z'y,$$

e a matriz de covariância de $\hat{\alpha}$ é

$$V(\hat{\alpha}) = \sigma^2(Z'Z)^{-1} = \sigma^2\Lambda^{-1}.$$

Assim, um autovalor pequeno de $X'X$ significa que a variância do correspondente coeficiente de regressão ortogonal será grande. Como

$$Z'Z = \sum_{i=1}^{p}\sum_{j=1}^{p} Z_i Z_j^{'} = \Lambda ,$$

frequentemente, o autovalor $\lambda_j$ é referido como a variância do j-ésimo componente principal. E a matriz de covariância dos coeficientes de regressão padronizados $\hat{\beta}$ é

$$V(\hat{\beta}) = V(T\hat{\alpha}) = T\Lambda^{-1}T'\alpha^2 .$$

Para a obtenção do estimador dos componentes principais, as variáveis independentes são consideradas em ordem decrescente de seus autovalores, isto é, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Os “s” últimos desses autovalores, com $s < p$, são considerados como sendo aproximadamente iguais a zero. As colunas da matriz correspondentes a esses autovalores próximos de zero, são excluídas da análise, e a regressão em componentes principais é, então, obtida pela aplicação do método dos mínimos quadrados aos componentes restantes. Isto é,

$$\hat{\alpha}_{PC} = B\hat{\alpha} ,$$

onde $b_1 = b_2 = \ldots = b_{p-s} = 1$ e $b_{p-s+1} = b_{p-s+2} = \ldots = b_p = 0$. Então, o estimador em componentes principais é

$$\hat{\alpha}_{PC} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{p-s} \\ --- \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

ou em termos de variáveis padronizadas

$$\hat{\beta}_{PC} = T\hat{\alpha}_{PC}$$
$$= \sum_{J=1}^{p-s} \lambda_j^{-1} t_j' X' y t_j .$$

De acordo com Montgomery e Peck (1992), em um estudo de simulação, Gunst e Mason em 1977, mostram que a regressão em componentes principais oferece considerável melhoria sobre os mínimos quadrados quando os dados são mal-condicionados.

## REFERÊNCIAS BIBLIOGRÁFICAS


AMARAL JÚNIOR, A. T.; CASALI, V. W. D.; CRUZ, C. D.; SILVA, D. J. H. da; SILVA, L. F. C. Estimativas de correlações fenotípicas, genotípicas e de ambiente entre sete caracteres morfoagronômicos em oito acessos de moranga. **Bragantia**, v. 53, n. 2, p. 163-166, 1994.

BELSLEY, D. A.; KUH, E.; WELSCH, R. E. **Regression Diagnostics: Identifying influential data and sources of collinearity**. New York: John Wiley & Sons, 1980. 292p.

CARVALHO, S. P. **Métodos alternativos de estimação de coeficientes de trilha e índices de seleção, sob multicolinearidade**. Viçosa, 1995. 163 p. Tese (Doutorado em Genética e Melhoramento). Universidade Federal de Viçosa, Viçosa.

CHATTERJEE, S.; PRICE, B. **Regression analysis by example**. New York: John Wiley & Sons, 1977. 228p.

COIMBRA, J. L. M.; BENIN, G.; VIEIRA, E. A.; OLIVEIRA, A. C. de; CARVALHO, F. I. F.; GUIDOLIN, A. F.; SOARES, A. P. Consequências da multicolinearidade sobre a análise de trilha em canola. **Ciência Rural**, v. 35, n. 2, p. 347-352, 2005.

CRUZ, C.D. **Programa Genes: Estatística experimental e matrizes.** 1ª ed. Viçosa: Editora UFV, 2006. 285p.

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** v. 2. 2. ed. rev. Viçosa: Editora UFV, 2006.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** v.1. 3. ed. Viçosa: Editora UFV, 2004. 480p.

DEWEY, D. R.; LU, K. H. A correlation and path coefficient analysis of components of crested wheatgrass seed production. **Agronomy Journal**, v.51, p.515-518, 1959.

DUNN, O. J.; CLARK, V. A. **Applied Statistics: Analysis of variance and regression**. New York: John Wiley & Sons, 1974. 386 p.

FALCONER, D.S. **Introdução à genética quantitativa.** Viçosa: Imprensa Universitária da UFV, 1987.

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**,v. 66, n. 4, p. 605-610, 2007.

GOES, T.; ARAÚJO, M. de; MARRA, R. Novas fronteiras tecnológicas da cana-de-açúcar no Brasil. **Revista de Política Agrícola**, n. 1, p. 50-59, 2009.

GUNST, R. F.; MASON, R. L. Advantages of examining multicollinearities in regression analysis. **Biometrics**. v. 33, p. 249-260, 1977.

GUJARATI, D. N. **Econometria básica**. 3. ed. São Paulo: Pearson Makron Books, 2000.

HOCKING, R. R. The analysis and selection of variables in linear regression. **Biometrics**. v. 32, p. 1-39, 1976.

HOERL, A. E. Optimum solution of many variable equations. **Chemical Engineering Progress.** v. 55, p. 69-78, 1959.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**. v. 12, n. 1, p. 55-67, 1970a.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Applications to nonorthogonal problems. **Technometrics**. v. 12, n. 1, p. 69-82, 1970b.

HOERL, R. W. Ridge analysis 25 years later. **The American Statistician**. v. 39, n. 3, p. 186-192, 1985.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. New Jersey: Prentice-Hall, 3. ed., 1992. 607p.

KANG, M.S.; MILLER, J.D.; TAI, P.Y.P. Genetic and phenotypic path analysis and heritability in sugarcane. **Crop Science**, v. 23, n. 4, p.643-647, 1983.

KENDALL, M. G. **A course in multivariate analysis**. London: Griffing, 1957. 185p.

KUTNER, M. H; NACHTSHEIM, C. J.; NETER, J.; LI, W. **Applied linear models**. Boston: Mcgraw-Hill Irwin, 5. ed., 2005. 1396p.

LAWSON, C. L.; HANSON, R. J. **Solving least square problems**. Prentice-Hal: Englewood Cliffs, 1974. 340p.

MANDEL, J. Use of the singular value decomposition in regression analisys. **The American Statistician**, v. 36, n. 1, p. 15-24, 1982.

MARQUARDT, D. W. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. **Technometrics**. v. 13, n.3, p. 591-612, 1970.

MIRANDA, J. E. C. de; COSTA, C. P. da; CRUZ, C. D. Correlações genotípica, fenotípica e de ambiente entre caracteres de fruto e planta de pimentão. **Revista Brasileira de Genética**, v. 11, n. 2, p. 457-468, 1988.

MONTGOMERY, D.C.; PECK, E.A. **Introduction to Linear Regression Analysis**. 2. ed., New York: John Wiley & Sons, 1992. 544p.

NETER, J.; WASSERMAN, W. **Applied linear statistical models: Regression, analysis of variance, and experimental designs. 3. ed.**. Homewood: Richard D. Irwin, 1990, 1181p.

NEWELL, G. J.; LEE, B. Ridge regression: An alternative to multiple linear regression for highly correlated data. **Journal of food science**. v. 46, p. 968-969, 1981.

PELUZIO, J. M.; SEDIYAMA, C. S.; SEDIYAMA, T.; REIS, M. S. Correlações fenotípicas, genotípicas e de ambiente entre alguns caracteres de soja, em Pedro Afonso, Tocantins. **Revista Ceres**, v. 45, n. 259, p. 303-308, 1998.

PETERNELLI, L. A. Comunicação pessoal. 2009.

SILVEY, S. D. Multicollinearity and imprecise estimation. **Journal of the Royal Statistical Society**. Series B, v. 31, n. 3, p. 589-652, 1969.

WRIGHT, S. Correlation and causation. **Journal of Agricultural Research**, v. 20, n. 7, p. 557-585, 1921.

## CAPÍTULO 1

## ANÁLISE DE TRILHA PARA COMPONENTES DO RENDIMENTO NA SELEÇÃO DE FAMÍLIAS DE CANA-DE-AÇÚCAR

### RESUMO

A análise de trilha é um recurso que o melhorista dispõe para entender as causas envolvidas nas associações entre caracteres e decompor a correlação existente em efeitos diretos e indiretos, através de uma variável principal e das suas variáveis explicativas. O presente trabalho foi proposto para quantificar os efeitos diretos e indiretos, por meio da análise de trilha, utilizando valores fenotípicos e genotípicos dos componentes de produção – número de colmos por parcela, diâmetro médio de colmos e comprimento médio de colmos – sobre a produtividade de colmos por hectare em cana-de-açúcar, nas fases de cana-planta e cana-soca, em etapa inicial de seleção do programa de melhoramento da cana-de-açúcar no estado de Minas Gerais. Dados de cana-planta e cana-soca da fase inicial de seleção foram obtidos de dois experimentos. Cada experimento foi constituído de vinte e duas famílias de cana-de-açúcar, e conduzido no Centro de Experimentação em Cana-de-açúcar (CECA), da Universidade Federal de Viçosa, no delineamento experimental em blocos casualizados com cinco repetições. O plantio foi realizado em abril de 2007. Foram avaliados os caracteres tonelada de colmos por hectare (*TCH*), variável principal; e seus componentes de produção, variáveis explicativas: comprimento médio de colmos (*CC*), diâmetro médio de colmos (*DC*) e número de colmos (*NC*). Os coeficientes de determinação foram elevados em todas as análises de trilha, indicando que os componentes avaliados explicam grande parte da variação existente na produção de colmos. Pela análise dos efeitos diretos fenotípicos e genotípicos, *NC* foi a variável que melhor se correlacionou com *TCH*, em ambos os experimentos e estágios, demonstrando a possibilidade de obtenção de ganhos significativos por meio da seleção indireta para *TCH* via

*NC*. A avaliação das relações de causa e efeito entre os componentes de produção em cana-de-açúcar possibilitou verificar que houve variação entre os experimentos, o que provavelmente se deve à origem diferenciada das famílias avaliadas.

# 1. INTRODUÇÃO

Nos programas de melhoramento genético de cada cultura, a estimativa da correlação entre as variáveis é importante quando se deseja realizar a seleção simultânea entre as características ou quando o caráter de interesse apresenta baixa herdabilidade ou difícil mensuração ou identificação (Marchezan, et al., 2005).

O conhecimento das relações existentes entre variáveis é de grande importância, visto que a obtenção de ganhos genéticos e a definição dos melhores genótipos são, muitas vezes, dirigidas a um conjunto de variáveis agronômicas e comerciais (Ferreira, 2007).

O grau de associação linear entre duas variáveis quaisquer pode ser definido pela correlação de Pearson. Robinson e Cockerhan (1965) mencionam que as correlações, juntamente com as variâncias, são os parâmetros que mais interessam no melhoramento de uma cultura.

Segundo Falconer (1987), nos estudos genéticos, é necessário distinguir duas causas de correlação entre características: a genética e a de ambiente. A causa de correlação genética é, principalmente, o pleiotropismo – propriedade pela qual um gene condiciona mais de um caráter simultaneamente –, embora ligações gênicas sejam uma causa de correlação transitória. O ambiente é uma causa de correlação, pela qual duas características são influenciadas pelas mesmas diferenças de condições de ambientes. Ainda de acordo com este mesmo autor, a correlação fenotípica é definida como a associação entre duas variáveis que pode ser observada diretamente.

As relações existentes entre os caracteres são, em geral, avaliadas por meio das correlações fenotípicas, genotípicas e ambientais. De acordo com Cruz et al. (2004), a correlação fenotípica tem causas genéticas e ambientais, porém, somente as genéticas envolvem uma associação de natureza herdável, podendo, portanto, ser utilizada para orientar programas de melhoramento. Assim, em estudos genéticos é indispensável distinguir e quantificar o grau de associação genética e ambiental entre os caracteres.

É de grande utilidade em programas de melhoramento vegetal a estimação dos coeficientes de correlações genotípica, fenotípica e de ambiente dos caracteres intimamente ligados à produção de grãos ou frutos, uma vez que esse é um caráter complexo, governado por vários genes e quase sempre de baixa herdabilidade (Falconer, 1987).

Em alguns casos, a seleção com base em uma característica de fácil avaliação que está altamente correlacionada com a variável de difícil seleção pode levar a progressos mais rápidos. Dessa forma, utiliza-se a correlação entre caracteres, pois, através do conhecimento da magnitude do desempenho de uma característica, pode-se avaliar a influência sobre a outra característica.

Todavia, podem ocorrer alguns equívocos nas estratégias de seleção das características avaliadas a partir da quantificação da magnitude das correlações entre as variáveis, pois um alto ou baixo coeficiente de correlação entre dois caracteres pode ser resultado do efeito de um terceiro sobre eles, ou de um grupo de caracteres (Cruz et al., 2004). Portanto, para o melhor entendimento dos fenômenos de associação entre as variáveis, o estudo da análise de trilha é indispensável no melhoramento.

A análise de trilha foi desenvolvida pelo geneticista Sewall Wright em 1918-1921 para explicar as relações causais em genética de população (Johnson e Wichern, 1992). Seu objetivo é fornecer explicações plausíveis de correlações observadas pela construção de modelos de relação de causa e efeito entre as variáveis. De acordo com Cruz (2006), a análise de trilha, apesar de envolver princípios de regressão, é, em essência, um estudo da decomposição do coeficiente de correlação, permitindo avaliar se a relação entre duas variáveis é de causa e efeito ou se é determinada pela influência de outra(s) variável(is).

De acordo com Li (1956), a análise de trilha é um método de análise multivariada apropriado para lidar com um sistema “fechado” de variáveis linearmente relacionadas. Caso estas variáveis não se relacionem linearmente, faz-se necessária uma transformação de escalas para que o método possa ser aplicado. Por sistema linear fechado, entende-se que cada

variável pode ser tanto uma combinação linear de algumas outras variáveis, como pode ser um dos fatores básicos, e nesse caso pode ainda ser correlacionada ou não a outros fatores do sistema.

Embora seja a correlação uma característica intrínseca a dois caracteres em dada condição experimental, sua decomposição depende do conjunto de caracteres estudados. Esses caracteres, normalmente, são avaliados pelo conhecimento prévio do pesquisador de sua importância e de possíveis inter-relações expressas em “diagramas de trilha” (Cruz et al., 2004). A construção gráfica de esquema causal possibilita a obtenção de um conjunto de equações simultâneas. De acordo com Li (1956), o diagrama causal é baseado em um conhecimento a priori de relações causais, ou em uma hipótese que o investigador escolhe para que seja testada.

Dewey e Lu (1959), em seus estudos sobre correlações e coeficientes de trilha, exemplificaram o uso deste último método e evidenciaram sua utilidade na análise dos coeficientes de correlação, quando se estuda o inter-relacionamento de caracteres agronômicos desejáveis.

Atualmente, a análise de trilha é amplamente utilizada por melhoristas de plantas (amendoim, por Santos et al., 2000; feijão, por Kurek et al., 2001; arroz, por Marchezan et al., 2005; algodão, por Hoogerheide et al., 2007; trigo, por Gondim et al., 2008). Estudos sobre relações entre variáveis importantes para a cultura da cana, via análise de trilha, também são encontrados na literatura (Kang et al., 1983; Reddy e Reddi, 1986; Sukhchain et al., 1997; Barbosa et al., 2002; Ferreira et al., 2007; Silva et al., 2009).

O emprego da análise de trilha nos programas de melhoramento da cana-de-açúcar visa apontar as características mais adequadas para que seja feita uma seleção indireta dos genótipos mais produtivos, uma vez que quantificar a produção destes genótipos é um trabalho bastante demorado devido ao grande número de genótipos avaliados nas etapas iniciais.

Segundo Reddy e Reddi (1986), o rendimento de cana constitui uma característica complexa influenciada por vários caracteres inter-relacionados. A interdependência entre os caracteres frequentemente influencia a relação direta com o rendimento e, como resultado, a informação baseada em

coeficientes de correlação se torna não confiável. Mas a análise dos coeficientes de trilha permite a decomposição dos coeficientes de correlação em efeitos diretos e indiretos e fornece uma relação mais prática dos caracteres, ajudando na identificação de componentes de grande efeito.

Assim, a análise de trilha aparece como um recurso bastante útil para otimizar o processo de identificação de genótipos (ou indivíduos) promissores nos programas de melhoramento genético com base em características indiretas. Do mesmo modo, o estudo de formas mais eficientes de se realizar tais análises são de suma importância dentro dos objetivos dos programas de melhoramento de plantas, em especial, o da cana-de-açúcar.

No entanto, estudos sobre as relações entre variáveis importantes para a cultura da cana, via análise de trilha, ainda são necessários, visto que diferentes estruturas genéticas populacionais têm sido consideradas (Ferreira, 2007).

Dessa forma, o objetivo deste trabalho foi quantificar os efeitos diretos e indiretos, por meio da análise de trilha, utilizando valores fenotípicos e genotípicos dos componentes de produção – número de colmos por parcela, diâmetro médio de colmos e comprimento médio de colmos – sobre a produtividade de colmos por hectare em cana-de-açúcar, nas fases de cana-planta e cana-soca, em etapa inicial de seleção do programa de melhoramento da cana-de-açúcar, no estado de Minas Gerais, favorecendo, assim, o processo de seleção indireta de genótipos mais produtivos

## 2. MATERIAL E MÉTODOS

Foram utilizados dados de dois experimentos de cana-de-açúcar. Cada experimento foi constituído de vinte e duas famílias resultantes de vinte e dois cruzamentos biparentais de diferentes genótipos de cana-de-açúcar. Esses cruzamentos foram realizados em 2006, na Estação de Floração e Cruzamentos, localizada na Serra do Ouro, município de Murici – AL, situada à latitude 9º13' S, longitude 35º50'W e a 450-500m de altitude. As plântulas obtidas de cada família foram transplantadas, conforme metodologia descrita por Barbosa e Silveira (2000), em abril de 2007, no Centro de Experimentação em cana-de-açúcar (CECA), localizado em Oratórios – MG (latitude 20º25'S, longitude 42º48'W e 494m de altitude), pertencente à Universidade Federal de Viçosa.

Cada parcela foi constituída por dois sulcos espaçados de 1,40 m, sendo cada sulco composto por dez plantas eqüidistantes a 0,5 m. A adubação dos experimentos foi realizada de acordo com aquela recomendada para a cultura (Korndörfer et al., 1999). Foi utilizado, nos experimentos, o delineamento em blocos casualizados com cinco repetições.

Em junho de 2008 e julho de 2009, no estágio de cana-planta e cana-soca, respectivamente, da fase inicial de seleção (Fase T1), foram avaliados em cada parcela, os caracteres: a) comprimento médio de colmos (*CC*) em metros, mensurando-se um colmo (amostrado aleatoriamente) de cada touceira, desde a base do colmo até o primeiro dewlap visível; b) diâmetro médio de colmos (*DC*) em milímetros, com a amostragem feita no quinto entrenó, contado da base do colmo para o ápice, mensurando-se um colmo (amostrado aleatoriamente) de cada touceira com paquímetro; c) número de colmos por parcela (*NC*); d) tonelada de colmos por hectare ($TCH_{direto}$), obtido de forma direta, pela expressão

$$TCH_{direto} = (\text{peso total da parcela} \times 10)/tp \,,$$

em que *tp* é o tamanho da parcela em m², sendo neste trabalho igual a 14 m².

Os componentes de produção – número de colmos por parcela (*NC*), diâmetro médio de colmos (*DC*) e comprimento médio de colmos (*CC*) –

explicam a produtividade de colmos por hectare (*TCH*) em cana-de-açúcar, de forma indireta, pela expressão apresentada por Ferreira et al. (2007):

$$TCH_{indireto} = d \times \pi \times NC \times CC \times \left(\frac{DC}{2}\right)^2 \times \frac{1}{100 \times tp}$$

onde *CC* e *DC* são expressas em centímetros e *d* é a densidade específica do colmo, g $cm^{-3}$. Chang e Milligan (1992) sugeriram utilizar para *d* o valor 1,0 g $cm^{-3}$.

Como a relação entre as variáveis explicativas *NC*, *DC* e *CC* e a variável principal *TCH* é estruturalmente multiplicativa, foi estabelecida a transformação dos dados para a escala logarítmica, de modo que fosse obtida a determinação completa do seguinte modelo aditivo de regressão linear múltipla (Ferreira et al., 2007):

$$\log TCH = p_{NC} \log(NC) + p_{DC} \log(DC) + p_{AC} \log(CC) + p_{\varepsilon}\mu$$

em que: $p_{NC}$, $p_{DC}$, $p_{CC}$ são as medidas dos efeitos diretos (ou coeficientes de trilha) dos componentes de produção *NC*, *DC* e *CC*, respectivamente, sobre a variável principal *TCH*; $p_{\varepsilon}$ é o efeito direto de outras variáveis, não consideradas no modelo, sobre a variável principal; $\mu$ é o erro padronizado associado ao modelo; *NC*, *DC*, *CC* e *TCH* são os valores observados padronizados.

As análises de trilha foram realizadas conforme diagrama causal apresentado na Figura 1, no qual a produção de colmos é determinada pelos seus componentes, por meio dos efeitos diretos e indiretos de cada um desses componentes.

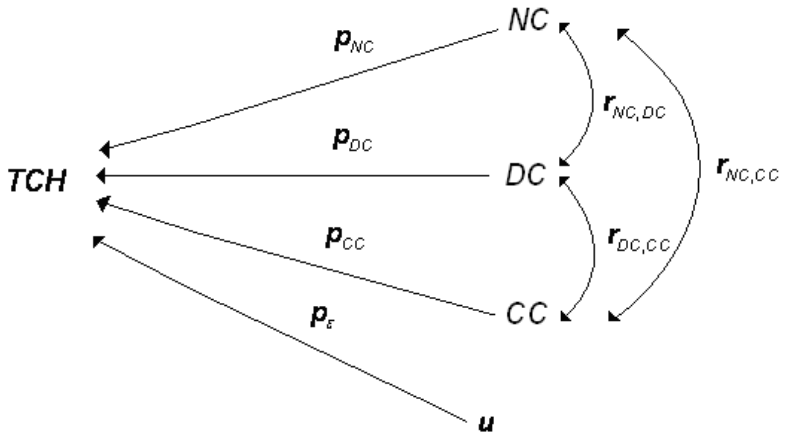


**Figura 1.** Diagrama causal ilustrando os efeitos diretos e indiretos das variáveis explicativas: número de colmos (*NC*) por parcela, diâmetro médio de colmos (*DC*) e comprimento médio de colmos (*CC*) sobre a variável principal tonelada de colmos por hectare (*TCH*). As setas em dupla direção indicam a associação mútua entre duas variáveis, determinada pelo coeficiente de correlação $(r_{ij})$ e as setas em única direção representam o efeito direto, determinado pelo coeficiente de trilha $(p_i)$.

Pelo desdobramento da correlação em efeitos diretos e indiretos, os coeficientes de trilha foram obtidos pela solução simultânea das seguintes equações:

$$r_{TCH\,NC} = p_{NC} + p_{DC}r_{NC\,DC} + p_{AC}r_{NC\,AC}$$
$$r_{TCH\,DC} = p_{NC}r_{DC\,NC} + p_{DC} + p_{AC}r_{DC\,AC}$$
$$r_{TCH\,AC} = p_{NC}r_{AC\,NC} + p_{DC}r_{AC\,DC} + p_{AC}$$

em que:

$r_{yi}$ : correlação entre a variável principal (*y*) e a i-ésima variável explicativa;

$p_i$ : medida do efeito direto da variável *i* sobre a variável principal;

$p_j r_{ij}$ : medida do efeito indireto da variável *i*, via variável *j*, sobre a variável principal.

O coeficiente de determinação foi calculado pela seguinte expressão:

$$R^2 = p_{NC}^2 + p_{DC}^2 + p_{AC}^2 + 2p_{NC}p_{DC}r_{NC\,DC} + 2p_{NC}p_{AC}r_{NC\,AC} + 2p_{DC}p_{AC}r_{DC\,AC}$$

O efeito residual sobre a variável principal foi estimado por meio de:

$$p_{\varepsilon} = \sqrt{1 - R^2}$$

Quanto à aplicação de testes de significância para as correlações, a literatura apresenta vários trabalhos (Barbosa et al., 2002, Ferreira, et al. 2007; Hoogerheide et al., 2007; Silva et al., 2009). Segundo Vencovsky e Barriga (1992) quando a correlação é fenotípica, a significância pode ser verificada por meio de teste t usual, conforme apresentado em diversos livros de Estatística (Exemplo: Dunn e Clark, 1974). No entanto, quando o coeficiente de correlação é estimado a partir de componentes de variância e da covariância, como acontece no caso da correlação genotípica entre caracteres, a fórmula usual para o teste t não mais se aplica. Assim, neste estudo, as significâncias das correlações fenotípicas e genotípicas foram avaliadas, respectivamente, conforme Dunn e Clark (1974) e Vencovsky e Barriga (1992).

Para que a avaliação do grau de associação entre diferentes caracteres de importância agronômica tenha uma estimativa confiável em termos biológicos, é fundamental identificar e quantificar o grau de multicolinearidade entre as variáveis estudadas. Assim, antes de realizar a análise de trilha, foi feito o diagnóstico de multicolinearidade nas matrizes de correlações genotípicas e fenotípicas.

Para a classificação da multicolinearidade foi realizado o exame do número de condição (*NC*) da matriz *X'X*, definido como a razão entre o maior e o menor autovalor dessa matriz. De acordo com Montgomery e Peck (1992), se o número de condição é menor que 100, a multicolinearidade não é um problema sério; para um valor de *NC* entre 100 e 1000 implica em multicolinearidade moderada a forte; se *NC* for maior que 1000, constitui indício de multicolinearidade severa.

Todas as análises foram efetuadas com o auxílio do programa computacional Genes (Cruz, 2009).

## 3. RESULTADOS E DISCUSSÃO

Os resultados da análise de variância de cada caráter quanto aos quadrados médios e coeficientes de variação, para cada experimento nos estágios de cana-planta e cana-soca, estão apresentados na Tabela 1.

**Tabela 1**. Resumo das análises de variância e estimativas dos coeficientes de variação experimental e das médias, relativas a quatro caracteres de cana-de-açúcar para o experimento 1 e 2 nas fases de cana-planta e cana-soca.

| | | Variável *TCH* | | | |
|---|---|---|---|---|---|
| | | Experimento 1 | | Experimento 2 | |
| | | Cana-planta | Cana-soca | Cana-planta | Cana-soca |
| FV | GL | QM | | | |
| Tratamentos | 21 | 1101,799** | 1683,9408** | 1790,3471** | 4321,007** |
| Blocos | 4 | 1080,368 | 575,7840 | 2296,5703 | 1192,6654 |
| Resíduo | 84 | 192,8171 | 538,5915 | 188,6848 | 528,8105 |
| Média | | 73,9188 | 110,5870 | 62,0286 | 107,1240 |
| CV(%) | | 18,79 | 20,99 | 22,15 | 21,47 |
| h² | | 82,50 | 68,01 | 89,46 | 87,76 |
| | | Variável *NC* | | | |
| | | Experimento 1 | | Experimento 2 | |
| | | Cana-planta | Cana-soca | Cana-planta | Cana-soca |
| FV | GL | QM | | | |
| Tratamentos | 21 | 1031,781** | 1137,1844** | 1727,1048** | 2414,0455** |
| Blocos | 4 | 1577,395 | 1010,4773 | 2130,9773 | 1359,7955 |
| Resíduo | 84 | 226,2859 | 434,9820 | 234,8677 | 376,8907 |
| Média | | 94,9909 | 114,4545 | 76,0 | 105,8636 |
| CV(%) | | 15,84 | 18,22 | 20,17 | 18,34 |
| h² | | 78,09 | 61,75 | 86,40 | 84,39 |
| | | Variável *DC* | | | |
| | | Experimento 1 | | Experimento 2 | |
| | | Cana-planta | Cana-soca | Cana-planta | Cana-soca |
| FV | GL | QM | | | |
| Tratamentos | 21 | 0,1111** | 0,0663** | 0,1934** | 0,0851** |
| Blocos | 4 | 0,0064 | 0,0140 | 0,0419 | 0,0801 |
| Resíduo | 84 | 0,0145 | 0,0146 | 0,0147 | 0,0190 |
| Média | | 2,3777 | 2,5276 | 2,385 | 2,5282 |
| CV(%) | | 5,04 | 4,78 | 5,08 | 5,46 |
| h² | | 87,08 | 77,98 | 92,40 | 77,66 |
| | | Variável *CC* | | | |
| | | Experimento 1 | | Experimento 2 | |
| | | Cana-planta | Cana-soca | Cana-planta | Cana-soca |
| FV | GL | QM | | | |
| Tratamentos | 21 | 639,257** | 1182,5637** | 2246,0107** | 2670,8563** |
| Blocos | 4 | 2485,869 | 269,3480 | 2670,9983 | 695,6063 |
| Resíduo | 84 | 180,7348 | 218,2783 | 248,0905 | 293,8546 |
| Média | | 199,0869 | 253,4674 | 201,1296 | 258,0306 |
| CV(%) | | 6,75 | 5,83 | 7,83 | 6,64 |
| h² | | 71,73 | 81,54 | 88,95 | 90,00 |

** Significativo a 1% pelo teste F.
*TCH*: Tonelada de colmos por hectare; *NC*: Número de colmos; *DC*: Diâmetro médio de colmos; *CC*: Comprimento médio de colmos.

Ocorreram diferenças significativas entre famílias em relação a todos os caracteres estudados. Houve boa precisão experimental em ambos os experimentos, pois a maioria dos caracteres apresentaram coeficiente de variação (CV) inferior a 20%), conforme critério de classificação de Gomes (1987).

Com base nos resultados das análises de variância, foram obtidas as estimativas do coeficiente de herdabilidade ($h^2$) no sentido amplo, em nível de parcela experimental. Os caracteres apresentaram $h^2$ relativamente elevados, acima de 60%, sendo todos considerados favoráveis para a seleção na cultura da cana-de-açúcar.

Pelo exame do número de condição, dado pela razão entre o maior e o menor autovalor da matriz *X'X*, foi possível identificar o grau de multicolinearidade presente. Assim, em ambos os experimentos e estágios, a multicolinearidade pode ser classificada como fraca, uma vez que o número de condição encontrado para cada matriz foi inferior a 100.

As estimativas dos coeficientes de correlação genotípica e fenotípica, entre os caracteres em estudo, para os estágios de cana-planta e cana-soca do experimento 1, estão apresentadas na Tabela 2.

**Tabela 2.** Estimativas dos coeficientes de correlação fenotípica e (genotípica) entre as variáveis: número de colmos (*NC*), diâmetro médio de colmos (*DC*), comprimento médio de colmos (*CC*) e tonelada de cana por hectare (*TCH*), avaliadas em famílias de cana-de-açúcar referentes ao experimento 1.

| Variáveis | Cana-planta | | | Cana-soca | | |
|---|---|---|---|---|---|---|
| | *DC* | *CC* | *TCH* | *DC* | *CC* | *TCH* |
| *NC* | 0,3199<br>(0,4583)* | 0,3004<br>(0,2798) | 0,8452**<br>(0,8587)** | 0,0844<br>(0,2613) | 0,2905<br>(0,3582) | 0,8063**<br>(0,7595) |
| *DC* | | 0,4191<br>(0,4493) | 0,5874**<br>(0,6962)** | | 0,4974*<br>(0,5643)* | 0,5756**<br>(0,8167)** |
| *CC* | | | 0,656**<br>(0,6941)* | | | 0,4748*<br>(0,534)* |

**,* Significativo a 1% e a 5% pelo teste t.

Embora a correlação genotípica entre *TCH* e *NC* (0,7595) seja relativamente alta, esta não foi significativa para a fase de cana-soca. Esse resultado se mostra semelhante ao encontrado por Ferreira et al. (2007), em que obtiveram correlações genotípicas não significativas entre *TCH* e *NC* e *TCH* e *CC* ao analisarem dados dessa mesma fase.

A significância da correlação genotípica foi verificada neste trabalho pela expressão $t = r_g \Big/ \sqrt{v\hat{a}r(r_g)}$, assim, quando a variância da correlação genotípica entre dois caracteres foi alta, o valor encontrado para t foi pequeno, implicando num baixo poder do teste, o que poderia explicar a não significância da correlação genotípica entre *TCH* e *NC*.

Observa-se que houve boa concordância de sinais nas correlações fenotípicas e genotípicas, para os dois estágios. Quanto às magnitudes, houve ligeira tendência das correlações genotípicas superarem as fenotípicas. Assim, é possível deduzir que os componentes genotípicos têm maior influência na determinação das correlações que os componentes de ambiente. No entanto, os valores encontrados para a correlação genotípica podem estar superestimados, uma vez que esta é proveniente de cálculos indiretos a partir de componentes da variância e da covariância.

Kang et al. (1983) também encontraram correlações fenotípicas e genotípicas com magnitudes próximas, permitindo concluir que a variância e a covariância ambiental foram reduzidas a nível desprezível, isto é, a influência do ambiente sobre estas relações foi pequena.

A maioria das correlações fenotípicas e genotípicas entre os componentes de produção (*NC*, *DC*, *C*) com *TCH* foram altas ou moderadas e positivas, à semelhança dos resultados obtidos por Ferreira et al. (2007). As maiores estimativas das correlações fenotípicas e genotípicas foram para número de colmos (*NC*), no estágio de cana-planta (0,8452 e 0,8587, respectivamente) e no estágio de cana-soca (0,8063 e 0,7595, respectivamente). Considerando apenas os coeficientes de correlação, o componente de produção *NC* é, portanto, de principal importância na determinação de *TCH*.

A estimativa do coeficiente de correlação genotípica entre *DC* e *TCH*, no estágio de cana-soca (0,8167) foi superior à fase de cana-planta (0,6962), no entanto, o mesmo não ocorreu para os demais caracteres.

A Tabela 3 apresenta as estimativas das correlações genotípicas e fenotípicas, entre os caracteres em estudo, para os estágios de cana-planta e cana-soca do experimento 2.

**Tabela 3.** Estimativas dos coeficientes de correlação fenotípica e (genotípica) entre as variáveis: número de colmos (*NC*), diâmetro médio de colmos (*DC*), comprimento médio de colmos (*CC*) e tonelada de cana por hectare (*TCH*), avaliadas em famílias de cana-de-açúcar referentes ao experimento 2.

| | Cana-planta | | | Cana-soca | | |
|---|---|---|---|---|---|---|
| Variáveis | *DC* | *CC* | *TCH* | *DC* | *CC* | *TCH* |
| *NC* | 0,6702** (0,7812)** | 0,7014** (0,7633)** | 0,9434** (0,9550)** | 0,4648* (0,6149)** | 0,7032** (0,7458)** | 0,9393** (0,9473)** |
| *DC* | | 0,6879** (0,7146)** | 0,8414** (0,9103)** | | 0,6675** (0,7380)** | 0,6868** (0,8168)** |
| *CC* | | | 0,8230** (0,8547)** | | | 0,8312** (0,8632)** |

**,* Significativo a 1% e a 5% pelo teste t.

As estimativas dos coeficientes de correlação fenotípica e genotípica entre *TCH* e seus componentes número de colmos (*NC*), diâmetro médio de colmos (*DC*) e comprimento médio de colmos (*CC*) foram significativas a 1% pelo teste t. Resultado concordante foi encontrado no trabalho de Silva et al. (2009) para o estágio de cana-planta.

As estimativas dos coeficientes de correlação fenotípica e genotípica entre *TCH* e os seus componentes número de colmos (*NC*), diâmetro médio de colmos (*DC*) e comprimento médio de colmos (*CC*), no estágio de cana-planta, foram positivos e elevados (superiores a 0,80), sugerindo, inicialmente que um aumento em qualquer um desses componentes causaria aumento correspondente em *TCH*.

Assim, caso apenas os coeficientes de correlação anteriormente citados fossem considerados neste estudo, os três componentes avaliados

teriam praticamente a mesma importância ao determinar *TCH*, já que as estimativas apresentaram valores bastante próximos. Tal estudo está de acordo com aqueles apresentados por Silva et al. (2009). Observa-se ainda elevada correlação fenotípica e genotípica entre as características *NC* e *DC* (0,6702 e 0,7812, respectivamente), *NC* e *CC* (0,7014 e 0,7633, respectivamente) e *DC* e *CC* (0,6879 e 0,7146, respectivamente), conforme Tabela 3.

No estágio de cana-soca, as estimativas dos coeficientes de correlação fenotípica e genotípica se mostraram similares em magnitude em relação às estimativas da fase de cana-planta, havendo uma pequena redução nas correlações fenotípicas e genotípicas entre as variáveis *TCH* e *DC.*

No estudo de correlação, quanto mais variáveis são consideradas, mais complexas se tornam as associações indiretas entre elas. Dessa forma, o coeficiente de trilha fornece um meio eficaz de desdobramento das causas diretas e indiretas de associação, permitindo medir a importância relativa de cada fator causal (Dewey e Lu, 1959).

Assim, para o experimento 1, os coeficientes de correlação das variáveis explicativas *NC*, *DC* e *CC* sobre *TCH* foram decompostos em efeitos diretos e indiretos, como apresentado na Tabela 4.

**Tabela 4.** Análise de trilha fenotípica e genotípica dos componentes de produção número de colmos (*NC*), diâmetro médio de colmos (*DC*), comprimento médio de colmos (*CC*) sobre tonelada de cana por hectare (*TCH*) para o experimento 1.

| | Cana-planta | | Cana-soca | |
|---|---|---|---|---|
| Variáveis | Fenotípica | Genotípica | Fenotípica | Genotípica |
| *NC* | | | | |
| Efeito direto sobre *TCH* | 0,6650 | 0,6440 | 0,7636 | 0,6041 |
| Efeito indireto via *DC* | 0,0712 | 0,0977 | 0,0432 | 0,1839 |
| Efeito indireto via *CC* | 0,1090 | 0,1170 | -0,0005 | -0,0285 |
| Total | 0,8452 | 0,8587 | 0,8063 | 0,7595 |
| *DC* | | | | |
| Efeito direto sobre *TCH* | 0,2226 | 0,2132 | 0,5120 | 0,7037 |
| Efeito indireto via *NC* | 0,2127 | 0,2952 | 0,0644 | 0,1578 |
| Efeito indireto via *CC* | 0,1521 | 0,1879 | -0,0008 | -0,0449 |
| Total | 0,5874 | 0,6962 | 0,5756 | 0,8167 |
| *CC* | | | | |
| Efeito direto sobre *TCH* | 0,3630 | 0,4181 | -0,0017 | -0,0795 |
| Efeito indireto via *NC* | 0,1998 | 0,1802 | 0,2218 | 0,2164 |
| Efeito indireto via *DC* | 0,0933 | 0,0958 | 0,2547 | 0,3971 |
| Total | 0,6560 | 0,6941 | 0,4748 | 0,5340 |
| Coeficiente de determinação ($R^2$) | 0,9309 | 0,9916 | 0,9096 | 0,9911 |
| Efeito da variável residual | 0,2629 | 0,0914 | 0,3007 | 0,0945 |

Tanto nas correlações fenotípicas quanto genotípicas, os efeitos diretos e indiretos dos caracteres *NC*, *DC* e *CC* sobre *TCH* foram todos positivos no estágio de cana-planta. Os diretos foram superiores aos indiretos ao se analisar *NC* e *CC*. No entanto, para a variável *DC*, os efeitos diretos e indiretos apresentaram, praticamente, as mesmas magnitudes, refletindo os valores de correlações fenotípicas e genotípicas entre as variáveis *DC* e *NC* (0,3199 e 0,4583, respectivamente) e *DC* e *CC* (0,4191 e 0,4493, respectivamente) como apresentado na Tabela 1, fazendo com que houvesse fracionamento por igual do coeficiente de correlação entre os efeitos direto e indireto.

Embora as variáveis *DC* e *CC* tenham apresentado estimativas de correlação genotípica e fenotípica moderadas com *TCH*, os efeitos diretos foram baixos. Sendo o efeito direto da variável *DC* (0,2226) inferior ao valor de efeito da variável residual (0,2629), o que reduz sua importância em relação às variáveis *NC* e *CC*.

Verifica-se ainda pela Tabela 4 que o efeito indireto genotípico de *DC* via *NC* (0,2952) foi mais importante que o próprio efeito direto (0,2132) sobre *TCH*. Segundo Vencovsky e Barriga (1992), caracteres com alta correlação favorável, como é o caso da variável *DC* em cana-planta, com correlação genotípica igual a 0,6962, mas com baixo efeito direto, indicam que a melhor estratégia deverá ser a seleção simultânea de caracteres, com ênfase também nos caracteres cujos efeitos indiretos são significativos.

Na fase de cana-planta, destaca-se *NC* como a variável que mais contribuiu para explicar *TCH*. Esse resultado está de acordo com os obtidos por Silva et al. (2009) e Barbosa et al. (2002), que obtiveram altos efeitos diretos de *NC* sobre *TCH*. Os trabalhos de James (1971), Reddy e Reddi (1986), apresentaram contribuição semelhante de *NC* e *DC* sobre *TCH*, seguidos de *CC*. Sukhchain et al. (1997) também obtiveram altos efeitos diretos de *NC* sobre *TCH*, sugerindo a seleção de clones para elevação da produção de colmos com base nesta variável.

Os coeficientes de trilha fenotípicos (Tabela 4) explicaram bem as variações em *TCH*, como indica o alto valor do coeficiente de determinação do modelo ($R^2$ = 0,9309) e pelo efeito residual pequeno (0,2629). Tal resultado reflete a excelente contribuição das variáveis do modelo para a produção de colmos, sendo estas variáveis dadas, algumas vezes, no cálculo de *TCH* indireto. Melhor ainda foi a análise usando valores genotípicos, que explicou em 99,11% a variação em *TCH*.

Foram encontrados no estágio de cana-soca, efeitos diretos positivos e superiores aos efeitos indiretos entre as variáveis *TCH e NC* e *TCH e DC*, tanto para as correlações fenotípicas quanto genotípicas. Contudo, o efeito direto de *CC* sobre *TCH* foi negativo e inferior aos efeitos indiretos, evidenciando a baixa contribuição dessa variável para *TCH* nessa fase (Tabela 4). Possivelmente, a correlação moderada positiva entre *TCH* e *CC* está sendo causada pelos efeitos indiretos, via *NC* e *DC*.

O comportamento apresentado pela variável *CC* em relação a *TCH* é indicativo da ausência de causa e efeito entre essas variáveis. Segundo Cruz et al. (2004), nessa situação, o caráter independente não é o principal

determinante das alterações na variável principal, existindo outros fatores que poderão proporcionar maior impacto em termos de ganhos de seleção.

A variável *DC* se destacou pela maior estimativa de efeito direto genotípico sobre *TCH* (0,7037), o que pode ser explicado pelo aumento da correlação genotípica entre *DC* e *TCH*.

O componente de produção *NC* apresentou a maior estimativa de efeito direto fenotípico sobre *TCH* (0,7636), com aumento desse efeito direto em relação à fase anterior da cultura (Tabela 4), indicando que no estágio de cana-soca, esse componente teve maior contribuição para a produção de cana, seguido de *DC*. Verifica-se que em ambas as fases da cultura, o componente *NC* se manteve como o mais explicativo. Resultados semelhantes foram obtidos por Silva et al. (2009).

No trabalho apresentado por Kang et al. (1983) a decomposição dos coeficientes de correlação genotípica mostrou que diâmetro médio de colmos (*DC*) e número de colmos (*NC*) produziram igual contribuição para *TCH*, seguido de médio de colmos (*CC*). Para as correlações fenotípicas, os três componentes *NC*, *DC* e *CC*, apresentaram igual importância para a produção de cana-de-açúcar. De acordo com estes autores, ainda que do ponto de vista prático, os coeficientes de trilha genotípicos deveriam ser mais importantes para decidir sobre um critério de seleção mais eficaz.

Contudo, caracteres genotipicamente correlacionados, mas não fenotipicamente correlacionados podem não ser de valor prático na seleção, pois esta é geralmente baseada no fenótipo (Shukla et al., 1998). Assim, os dois tipos de correlação foram considerados neste estudo para facilitar a decisão sobre a eficiência de um critério de seleção indireta.

Assim como ocorreu na fase de cana-planta, tanto para as correlações fenotípicas quanto genotípicas, os coeficientes de determinação ($R^2$),0,9096 e 0,9911, respectivamente, e os baixos efeitos residuais, 0,3007 e 0,0945, respectivamente, representaram satisfatoriamente a contribuição de *NC*, *DC* e *CC* para *TCH* (Tabela 4).

As estimativas dos efeitos diretos e indiretos dos componentes de produção *NC*, *DC* e *CC* sobre *TCH*, do experimento 2, encontram-se na Tabela 5.

**Tabela 5.** Análise de trilha fenotípica e genotípica dos componentes de produção número de colmos (*NC*), diâmetro médio de colmos (*DC*), comprimento médio de colmos (*CC*) sobre tonelada de cana por hectare (*TCH*) para o experimento 2.

| | Cana-planta | | Cana-soca | |
|---|---|---|---|---|
| Variáveis | Fenotípica | Genotípica | Fenotípica | Genotípica |
| NC | | | | |
| Efeito direto sobre TCH | 0,6076 | 0,5211 | 0,7041 | 0,6397 |
| Efeito indireto via DC | 0,2051 | 0,2821 | 0,1133 | 0,1870 |
| Efeito indireto via CC | 0,1307 | 0,1519 | 0,1219 | 0,1205 |
| Total | 0,9434 | 0,9550 | 0,9393 | 0,9473 |
| DC | | | | |
| Efeito direto sobre TCH | 0,3060 | 0,3611 | 0,2439 | 0,3041 |
| Efeito indireto via NC | 0,4072 | 0,4071 | 0,3273 | 0,3934 |
| Efeito indireto via CC | 0,1282 | 0,1422 | 0,1157 | 0,1193 |
| Total | 0,8414 | 0,9103 | 0,6868 | 0,8168 |
| CC | | | | |
| Efeito direto sobre TCH | 0,1863 | 0,1989 | 0,1733 | 0,1616 |
| Efeito indireto via NC | 0,4262 | 0,3977 | 0,4951 | 0,4771 |
| Efeito indireto via DC | 0,2105 | 0,2580 | 0,1628 | 0,2245 |
| Total | 0,8230 | 0,8547 | 0,8312 | 0,8632 |
| Coeficiente de determinação ($R^2$) | 0,9841 | 0,9963 | 0,9729 | 0,9840 |
| Efeito da variável residual | 0,1262 | 0,0604 | 0,1647 | 0,0778 |

No estágio de cana-planta, os efeitos diretos e indiretos dos caracteres *NC*, *DC* e *CC* sobre *TCH* foram todos positivos, com efeitos diretos superiores aos indiretos para *NC*. Para as variáveis *DC* e *CC*, alguns dos efeitos indiretos superaram os diretos.

O componente de produção *NC* teve a maior estimativa de efeito direto sobre *TCH* tanto para a correlação fenotípica quanto genotípica (0,6076 e 0,5211, respectivamente). Assim, como ocorreu para o experimento 1, nessa fase da cana-de-açúcar, destaca-se *NC* como a variável que mais contribuiu para explicar *TCH*.

Os coeficientes de trilha fenotípicos e genotípicos explicaram satisfatoriamente as variações em *TCH*, como indica o alto valor do coeficiente de determinação ($R^2$ = 0,9841 e 0,9963, respectivamente) e o

efeito residual pequeno (0,1262 e 0,0604, respectivamente), refletindo a excelente contribuição das variáveis do modelo para a produção de colmos.

No estágio de cana-soca, os efeitos diretos dos componentes de produção *DC* e *CC* foram baixos, à semelhança da fase de cana-planta, evidenciando a baixa contribuição dessas variáveis para *TCH* (Tabela 5).

O componente de produção *NC* apresentou a maior estimativa de efeito direto sobre *TCH*, para a correlação fenotípica (0,7041) assim, como para a correlação genotípica (0,6397), à semelhança dos resultados obtidos anteriormente, indicando que em ambos os experimentos, constituídos de diferentes famílias, e estágios, esse componente teve maior contribuição para a produção de cana-de-açúcar. Verifica-se também que houve aumento desse efeito direto em relação à fase de cana-planta, o que pode ser justificado pela redução dos efeitos indiretos de *NC* via *DC* e *CC*.

Nessa fase da cana-de-açúcar, os coeficientes de trilha fenotípicos e genotípicos também explicaram grande parte das variações em *TCH*, como indica o alto valor do coeficiente de determinação ($R^2$ = 0,9729 e 0,9840, respectivamente) e o efeito residual pequeno (0,1647 e 0,0778, respectivamente), refletindo, mais uma vez, a excelente contribuição das variáveis do modelo para a produção de colmos.

O rendimento de cana constitui uma característica complexa influenciada por vários caracteres inter-relacionados, e a análise dos coeficientes de trilha ajuda na identificação dos componentes de grande efeito. Assim, a variável *NC*, de fácil mensuração, foi o componente que apresentou maior contribuição para a produção de cana-de-açúcar, com possibilidade de obtenção de ganhos significativos por meio da seleção indireta para *TCH* via *NC,* em ambos os experimentos e estágios.

Através dos dois experimentos foi possível avaliar diferentes famílias e comparar os resultados das análises de trilha. A variação apresentada entre os experimentos se deve, provavelmente, à origem diferenciada das famílias avaliadas.

# 4. CONCLUSÕES

Em ambos os experimentos e fases, houve maior influência dos componentes genotípicos, em relação aos fenotípicos, na determinação das correlações.

Os coeficientes de determinação indicaram que número de colmos (*NC*), diâmetro médio de colmos (*DC*) e comprimento médio de colmos (*CC*) explicaram grande parte da variação existente na produção de colmos.

A variável *NC* foi o componente que apresentou maior contribuição para a produção de cana-de-açúcar, com possibilidade de obtenção de ganhos significativos por meio da seleção indireta para *TCH* via *NC,* em ambos os experimentos e estágios.

A avaliação das relações de causa e efeito entre os componentes de produção em cana-de-açúcar possibilitou verificar que houve variação entre os experimentos, o que provavelmente se deve à origem diferenciada das famílias avaliadas. Assim, os resultados obtidos são pertinentes apenas para este estudo, havendo necessidade de avaliação de maior número de experimentos.

## REFERÊNCIAS BIBLIOGRÁFICAS


BAIN, L. J.; ENGELHARDT, M. **Introduction to probability and mathematical statistics**. 2. ed. California: Duxbury Press. 1992. 644 p.

BARBOSA, M. H. P.; BASTOS, I. T.; SILVEIRA, L. C. I.; OLIVEIRA, M. W. Análise de causa e efeito para produção de colmos e seus componentes na seleção de famílias de cana-de-açúcar. *In*: **8º Congresso Nacional da STAB**, 2002, Pernambuco. 8º Congresso Nacional da Sociedade dos Técnicos Açucareiros e Alcooleiros do Brasil. v.1, p. 366-370, 2002.

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, n. 3, p. 30-32, 2000.

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

CRUZ, C.D. **Programa Genes.** Viçosa: UFV. Versão 2009.

CRUZ, C.D. **Programa Genes: Estatística experimental e matrizes.** 1. ed. Viçosa: Editora UFV, 2006. 285p.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** v.1. 3. ed. Viçosa: Editora UFV, 2004. 480p.

DEWEY, D.R.; LU, K.H. A correlation and path coefficient analysis of components of crested wheatgrass seed production. **Agronomy Journal**, v.51, p.515-518, 1959.

DUNN, O. J.; CLARK, V. A. **Applied Statistics: Analysis of variance and regression**. New York: John Wiley & Sons, 1974. 386 p.

FALCONER, D.S. **Introdução à genética quantitativa.** Viçosa: Imprensa Universitária da UFV, 1987.

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**, v. 66, n. 4, p. 605-610, 2007.

GONDIM, T. C. de O.; ROCHA V.S.; SEDIYAMA, C.S.; MIRANDA, G.V. Análise de trilha para componentes do rendimento e caracteres agronômicos de trigo sob desfolha. **Pesquisa Agropecuária Brasileira**, v.43, n.4, p.487-493, 2008.

GOMES, F. P. Curso de estatística experimental. 12 ed. São Paulo: Nobel, 1987. 467p.

HOOGERHEIDE, E. S. S.; VENCOVSKY, R.; FARIAS, F. J. C.; FREIRE, E. C.; ARANTES, E. M. Correlações e análise de trilha de caracteres tecnológicos e a produtividade de fibra de algodão. **Pesquisa Agropecuária Brasileira**, v.42, n.10, p.1401-1405, 2007.

JAMES, N. I. Yield components in random and selected sugarcane populations. **Crop Science**, v. 11, n. 6, 906-908, 1971.

JOHNSON, R.A. & WICHERN, D. W. **Applied multivariate statistical analysis**. 2. ed. New Jersey, Prentice Hall, 1992, 607p.

KANG, M.S.; MILLER, J.D.; TAI, P.Y.P. Genetic and phenotypic path analysis and heritability in sugarcane. **Crop Science**, v. 23, n. 4, p.643-647, 1983.

KORNDÖRFER, G. H.; RIBEIRO, A. C.; ANDRADE L. A. B. **Sugestões de adubação para cana-de-açúcar**. In: RIBEIRO, A. C.; GUIMARÃES, P. T G.; ALVAREZ, V. H. Recomendações para o uso de corretivos e fertilizantes em Minas Gerais. 5ª aproximação. Viçosa, MG: Comissão de Fertilidade do Solo do Estado de Minas Gerais. p. 285-288, 1999.

KUREK, A.J.; CARVALHO, F.I.F. de; ASSMANN, I.C.; MARCHIORO, V.S.; CRUZ, P.J. Análise de trilha como critério de seleção indireta para rendimento de grãos em feijão. **Revista Brasileira de Agrociência**, v.7, n.1, p.29-32, 2001.

LI, C. C. **Path analysis: A primer**. Pacific Grove: Boxwood Press, 1975. 346p.

LI, C.C. The concept of path coefficient and its impact on population genetics. **Biometrics**, v.12, n.2, p. 190-210, 1956.

MARCHEZAN, E.; MARTIN, T. N.; SANTOS, F. M. dos; CAMARGO, E. R. **Análise de coeficiente de trilha para os componentes de produção em arroz.** Ciência Rural, v. 35, n.5, p.1027-1033, 2005.

MONTGOMERY, D.C.; PECK, E.A. **Introduction to Linear Regression Analysis**. 2. ed., New York: John Wiley & Sons, 1992. 504 p.

REDDY, C. R; REDDI, M. V. Degree of genetic determination, correlation and genotypic and phenotypic path analysis of cane and sugar yield in sugarcane. **Indian Journal of Genetics and Plant Breeding**, v. 46, n. 3, p. 550-557,1986.

ROBINSON, H. F.; C. C. COCKERHAN. Estimation y significado de los parametros genéticos. **Fitotecnia Latinoamericana**. Turrialba, v. 2, n. 1-2, p. 23-28, 1965.

SANTOS, R. C. dos; CARVALHO, L.P. de; SANTOS, V.F. dos. Análise de coeficiente de trilha para os componentes de produção em amendoim. **Ciência e Agrotecnologia**, v.24, n.1, p.13-16, 2000.

SHUKLA, S.; SINGH, K.; PUSHPENDRA. Correlation and path coefficient analysis of yield and its components in soybean (Glycine max L. Merrill.). **Soybean Genetics Newsletter**, Ames, v. 25, n.1, p. 67-70, 1998.

SILVA, F. L. da; PEDROZO, C. A.; BARBOSA, M. H. P.; RESENDE, M. D. V.; PETERNELLI, L. A.; Costa, P. M. de A. Análise de trilha para os componentes de produção de cana-de-açúcar via Blup. **Revista Ceres,** v. 56, n.3 , p. 308-314, 2009.

SUKHCHAIN; 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, n. 1, p. 109-114, 1997.

VENCOVSKY, R.; BARRIGA, P. **Genética biométrica no fitomelhoramento**. Ribeirão Preto: Revista Brasileira de Genética, 1992. 496 p.

# ANEXO 1

**Tabela 1.** Cruzamentos biparentais de cana-de-açúcar utilizados para estudo da análise de trilha do Experimento 1.

| Família | Cruzamento | | |
|---|---|---|---|
| 1 | SP81-3250 | x | LAICA 96-09 |
| 2 | IAC86-2210 | x | RB725053 |
| 3 | RB72454 | x | SP80-3280 |
| 4 | RB855113 | x | RB855156 |
| 5 | RB855584 | x | RB855046 |
| 6 | RB997810 | x | LAICA96-09 |
| 7 | RB92579 | x | SP80-1816 |
| 8 | IAC86-2210 | x | SP70-1143 |
| 9 | SP81-3250 | x | RB92579 |
| 10 | RB925345 | x | RB945957 |
| 11 | UFAL011569 | x | VAT90-61 |
| 12 | IAC86-2210 | x | CO 997 |
| 13 | RB855156 | x | SP71-1406 |
| 14 | RB867515 | x | RB855002 |
| 15 | RB825548 | x | RB92579 |
| 16 | RB998118 | x | Q 107 |
| 17 | L 60-14 | x | SP80-1842 |
| 18 | TUC71-7 | x | RB928064 |
| 19 | RB865513 | x | TUC71-7 |
| 20 | F 150 | x | RB855584 |
| 21 | RB855546 | x | RB92579 |
| 22 | SP91-1049 | x | VAT 90-212 |

# ANEXO 2

**Tabela 2.** Cruzamentos biparentais de cana-de-açúcar utilizados para estudo da análise de trilha do Experimento 2.

| Família | Cruzamento | | |
|---|---|---|---|
| 1 | RB825548 | x | RB935925 |
| 2 | RB855002 | x | RB945961 |
| 3 | RB725147 | x | RB92579 |
| 4 | RB935744 | x | RB806043 |
| 5 | RB845210 | x | RB845197 |
| 6 | RB735220 | x | SP77-5181 |
| 7 | RB825317 | x | RB835089 |
| 8 | RB935686 | x | RB815627 |
| 9 | RB92579 | x | RB855035 |
| 10 | RB92606 | x | RB805340 |
| 11 | RB935744 | x | RB845197 |
| 12 | RB865547 | x | RB855206 |
| 13 | H39-3633 | x | RB92606 |
| 14 | RB825237 | x | RB835486 |
| 15 | RB855546 | x | RB955970 |
| 16 | RB855063 | x | RB955970 |
| 17 | RB825237 | x | RB92579 |
| 18 | RB92606 | x | RB72199 |
| 19 | H64-1881 | x | RB75126 |
| 20 | RB855025 | x | RB931003 |
| 21 | RB855322 | x | RB835019 |
| 22 | RB825237 | x | H39-3633 |

# APÊNDICE

## PADRONIZAÇÃO DE VARIÁVEIS

A padronização de uma variável consiste em dividir o desvio de cada observação em relação à média, pelo respectivo desvio-padrão (Cruz et al., 2004). Segundo Li (1975), o processo de padronização faz das variáveis originais todas iguais em média (zero) e em variância (unidade), sendo desnecessárias as unidades físicas em que as variáveis são medidas.

De acordo ainda com este autor, a análise de trilha lida com as relações entre as variáveis padronizadas, em que os coeficientes de trilha são coeficientes de regressão parcial padronizados. Em essência, a análise de trilha é um estudo da decomposição do coeficiente de correlação (Cruz, 2006).

No entanto, é possível verificar que esse coeficiente de correlação permanece o mesmo se as variáveis são padronizadas ou não.

O coeficiente de correlação entre as variáveis X e Y é dado por

$$r_{XY} = \frac{C\hat{O}V(X,Y)}{\sqrt{\hat{V}(X)\hat{V}(Y)}}.$$

Aplicando-se a padronização nas variáveis X e Y, tem-se

$$r_{XY} = \frac{C\hat{O}V\left(\frac{X-\overline{X}}{\hat{\sigma}_X}, \frac{Y-\overline{Y}}{\hat{\sigma}_Y}\right)}{\sqrt{\hat{V}\left(\frac{X-\overline{X}}{\hat{\sigma}_X}\right)\hat{V}\left(\frac{Y-\overline{Y}}{\hat{\sigma}_Y}\right)}}.$$

Como $Var(aX+b) = a^2 Var(x)$ (Bain e Engelhardt, 1992),

$$r_{XY} = \frac{C\hat{O}V\left(\frac{X}{\hat{\sigma}_X} - \frac{\overline{X}}{\hat{\sigma}_X}, \frac{Y}{\hat{\sigma}_Y} - \frac{\overline{Y}}{\hat{\sigma}_Y}\right)}{\sqrt{\frac{1}{\hat{\sigma}_X^2\hat{\sigma}_Y^2}\hat{V}(X)\hat{V}(Y)}}.$$

Sendo $COV(aX, bY) = abCOV(X,Y)$ (Bain e Engelhardt, 1992),

$$r_{XY} = \frac{\frac{1}{\hat{\sigma}_X\hat{\sigma}_Y}C\hat{O}V(X,Y) - \frac{1}{\hat{\sigma}_X\hat{\sigma}_Y}C\hat{O}V(X,\overline{Y}) - \frac{1}{\hat{\sigma}_X\hat{\sigma}_Y}C\hat{O}V(\overline{X},Y) + \frac{1}{\hat{\sigma}_X\hat{\sigma}_Y}C\hat{O}V(\overline{X},\overline{Y})}{\sqrt{\frac{1}{\hat{\sigma}_X^2\hat{\sigma}_Y^2}\hat{V}(X)\hat{V}(Y)}}.$$

Sendo $COV(X + a, Y + b) = COV(X,Y)$ (Bain e Engelhardt, 1992),

$$r_{XY} = \frac{\frac{1}{\hat{\sigma}_X\hat{\sigma}_Y}C\hat{O}V(X,Y)}{\frac{1}{\hat{\sigma}_X\hat{\sigma}_Y}\sqrt{\hat{V}(X)\hat{V}(Y)}}.$$

$$\therefore$$

$$r_{XY} = \frac{C\hat{O}V(X,Y)}{\sqrt{\hat{V}(X)\hat{V}(Y)}}.$$

## CAPÍTULO 2

# ANÁLISE DE TRILHA SOB MULTICOLINEARIDADE EM CANA-DE-AÇÚCAR


## RESUMO

As técnicas estatísticas multivariadas têm sido regularmente aplicadas em diversas investigações científicas, gerando grande demanda por conhecimentos específicos, tanto a respeito da sua aplicação quanto das suas pressuposições ou limitações. Para que a avaliação do grau de associação entre diferentes caracteres tenha uma estimativa confiável, é fundamental identificar e quantificar o grau de multicolinearidade entre as variáveis estudadas. Os problemas causados pela multicolinearidade não são devidos simplesmente à sua presença, mas sim ao seu grau de manifestação. Com o objetivo de comparar o método baseado na regressão em crista e a exclusão de variáveis por componentes principais para a estimação dos coeficientes de trilha em presença de multicolinearidade, foram avaliados dados em cana-soca, obtidos do programa de melhoramento da cana-de-açúcar da Universidade Federal de Viçosa. O ensaio foi conduzido no delineamento em blocos casualizados, com oito cultivares e cinco repetições. Dez plantas por parcela foram amostradas para realização das análises das variáveis explicativas Brix (teor de sólidos solúveis), Pol (teor de sacarose aparente), pH (indica o grau de acidez), AR (açúcares redutores), ART (açúcares totais recuperáveis), Cu (cobre), Al (alumínio), Mg (magnésio), Ca (cálcio), K (potássio), Ácido aconítico, Compostos fenólicos, e da variável principal Cor ICUMSA. Esse material foi coletado e preparado conforme descrito por Santos (2008). A matriz de correlação obtida dos dados foi submetida a diferentes métodos para diagnóstico de multicolinearidade. Sob multicolinearidade severa, o método baseado na regressão em crista e a exclusão de variáveis por componentes principais apresentaram resultados semelhantes na estimação dos coeficientes de trilha, proporcionando sensível

redução na magnitude dos fatores de inflação da variância associados aos efeitos diretos e indiretos da análise de trilha. Assim, foi possível identificar, neste estudo, os caracteres alumínio (Al), potássio (K) e compostos fenólicos como aqueles que melhor explicam a Cor do caldo. Contudo, os demais caracteres devem ser levados em consideração devido a elevada correlação existente e a baixa magnitude do efeito direto, evidenciando a necessidade de seleção simultânea de caracteres, com ênfase também nos caracteres cujos efeitos indiretos são significativos. Para fins de melhoramento, a seleção indireta para Cor do caldo, por meio de índice de seleção envolvendo as variáveis Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos e Ácido aconítico é recomendável.

# 1. INTRODUÇÃO

Em problemas de regressão múltipla, é esperado encontrar dependências entre a variável resposta e as variáveis explicativas, no entanto, na maioria dos problemas de regressão, dependências também ocorrem entre as variáveis explicativas (Montgomery e Runger, 2008). Assim, segundo Neter et al. (1990) e Kutner et al. (2005), a multicolinearidade ocorre quando existe algum nível de inter-relação entre as variáveis independentes do modelo de regressão linear múltipla. Algumas vezes o termo multicolinearidade é utilizado apenas nos casos em que a correlação entre as variáveis é muito alta.

Para Vittinghoff et al. (2005), a multicolinearidade denota uma correlação entre variáveis explicativas, alta o suficiente para degradar substancialmente a precisão das estimativas dos coeficientes de regressão para algumas ou todas as variáveis explicativas correlacionadas.

Em muitos casos, quando é envolvido grande número de variáveis ou não há conhecimento prévio da associação entre elas, resultados inapropriados, gerados pela multicolinearidade, podem ser interpretados, levando a conclusões que não seriam as mais pertinentes (Cruz e Carneiro, 2006).

Como a análise de trilha é uma forma de estudo de regressão, baseada em matrizes de correlações, pode-se deduzir, por analogia, o virtual dano que a multicolinearidade causa ao processo de estimação dos coeficientes (Carvalho, 1995). Para obtenção dos efeitos diretos e indiretos da análise de trilha, é necessário, que a matriz $X'X$ esteja bem condicionada, contudo, os problemas de multicolinearidade podem torná-la singular, fazendo, consequentemente, com que as estimativas de mínimos quadrados não sejam confiáveis (Cruz e Carneiro, 2006).

Os métodos alternativos ao de mínimos quadrados, especificamente planejados para combater os problemas induzidos pela multicolinearidade, fornecem estimadores tendenciosos, mas, conforme Gunst e Mason (1977)

apresentam, em geral, melhor desempenho quando comparados aos estimadores de mínimos quadrados.

Um desses procedimentos consiste na exclusão de variáveis por meio da técnica de componentes principais (Montgomery e Peck, 1992), em que os componentes principais correspondentes aos autovalores próximos de zero são removidos da análise e o método dos mínimos quadrados é aplicado às variáveis restantes.

Quando a exclusão de variáveis não é de interesse do pesquisador, a análise de trilha é realizada com todas as variáveis, mas é adotado um procedimento equivalente ao da análise de regressão em crista, como recomendado por Carvalho (1995). Nesse método procura-se obter estimativas mais estáveis para os parâmetros do modelo, alterando ligeiramente o sistema de equações normais pela adição de uma constante à diagonal da matriz $X'X$, à semelhança do método de regressão em crista proposto por Hoerl e Kennard (1970a,b).

Para solucionar o problema é necessário fazer o diagnóstico de multicolinearidade dos dados antes de se realizar o processamento. Nesse diagnóstico são adotados alguns procedimentos básicos, considerando que o diagnóstico eficiente é aquele que, além de informar sobre a existência da multicolinearidade, apresenta o grau de severidade e identifica as variáveis envolvidas no problema (Montgomery e Peck, 1992).

Neste contexto, o estudo dos caracteres precursores da cor do açúcar através dos coeficientes de trilha permite entender as causas envolvidas nas associações entre esses caracteres e decompor a correlação existente em efeitos diretos e indiretos, através de uma variável principal, como a Cor do caldo de cana, e as variáveis explicativas Brix, Pol, pH, AR, ART, Cu, Al, Mg, Ca, K, Ácido aconítico e Compostos fenólicos.

De acordo com CLARKE & LEGENDRE (1999) a Cor do caldo de cana e do açúcar tem sua origem em vários compostos, como: flavonóides, compostos fenólicos, pigmentos e aqueles que reagem com os açúcares redutores, são os que mais afetam a Cor do caldo. Deste modo, o conhecimento desses compostos é imprescindível na qualidade do açúcar, já

que um dos indicadores utilizados na avaliação da qualidade do açúcar para exportação é a Cor. Assim, este trabalho tem por objetivo comparar o método baseado na regressão em crista e a eliminação de variáveis para a estimação dos coeficientes de trilha em presença de multicolinearidade, fazendo uso de dados em cana-soca, obtidos do programa de melhoramento da cana-de-açúcar da Universidade Federal de Viçosa, para avaliação das variáveis explicativas Brix, Pol, pH, AR, ART, Cu, Al, Mg, Ca, K, Ácido aconítico, Compostos fenólicos, sobre a variável principal Cor ICUMSA.

## 2. MATERIAL E MÉTODOS

Os dados foram obtidos de um experimento conduzido na área do Departamento de Fitotecnia da Universidade Federal de Viçosa, município de Viçosa, Minas Gerais. O plantio das mudas foi realizado em março de 2005 com o primeiro corte da cana-planta em maio de 2006. As avaliações ocorreram em cana-soca nos meses de abril e outubro de 2007.

Esse experimento foi delineado em blocos casualizados com oito cultivares de cana-de-açúcar, cinco repetições e parcelas de cinco sulcos com 10 m de comprimento. Para as análises tecnológicas foram amostradas dez colmos por parcela. Os tratamentos foram constituídos pelas cultivares RB72454, RB867515, RB835486, RB855156, SP80-1816, SP79-1011, RB855536 e RB92579.

Em abril e outubro de 2007, foram identificadas e coletadas aleatoriamente dez colmos por parcela, durante cinco dias consecutivos. Esse material foi coletado e preparado conforme descrito por Santos (2008). Após a extração do caldo, foram realizadas as análises tecnológicas, e as análises dos compostos orgânicos e inorgânicos, conforme Manual de controle químico da fabricação de açúcar – CTC (2001). Dentre as variáveis avaliadas foram utilizadas, nesse estudo, apenas as seguintes: análises tecnológicas (Brix, Pol, pH, AR, ART); compostos inorgânicos (Cu, Al, Mg, Ca, K); compostos orgânicos (Ácido aconítico, Compostos fenólicos) e Cor ICUMSA.

### 2.1. Obtenção dos dados e suas inter-relações

Nesta sessão são apresentadas, resumidamente, as variáveis utilizadas nas análises e suas inter-relações. Esta apresentação se faz necessária para que possa ser visualizada a forma de relacionamento dessas variáveis, ficando mais fácil discutir a possível existência de multicolinearidade entre essas variáveis.

• Brix (teor de sólidos solúveis) é o parâmetro mais utilizado na indústria do açúcar e do álcool. Este mede o índice de refração das soluções dissolvidas em uma solução açucarada fornecendo sua massa em porcentagem. Para a determinação do Brix na cana, é utilizada a seguinte equação:

$$\text{Brix\%cana} = \text{Brix do caldo} \times (1 - 0{,}01) \times C,$$

onde

C é o coeficiente que representa a transformação do caldo extraído em todo o caldo absoluto, dado por C = (1,0313 – 0,00575 x Fibra). A fibra é a matéria seca insolúvel na água, contida na cana-de-açúcar. Esta é determinada em função do Brix do caldo extraído da prensa hidráulica, peso de bagaço úmido (PBU) e peso de bagaço seco (PBS), conforme Fernandes (2003), por meio da expressão Fibra = (0,08 x PBU) + 0,876.

• Pol (teor de sacarose aparente) representa a porcentagem aparente de sacarose contida numa solução de açúcares. A pol na cana é obtida em função da Pol no caldo extraído, multiplicado pela fibra e pelo coeficiente “C” que transforma a Pol%caldo extraído em Pol%cana.

$$\text{Pol\%cana} = \text{Pol\%caldo} \times (1 - 0{,}01 \times \text{Fibra}) \times C,$$

em que

Pol%caldo = (1,0078 x leitura sacarimétrica + 0,0444) x (0,2607 – 0,009882 x Brix).

• Açúcares redutores (AR) são os principais precursores da cor mais escura do açúcar no processo industrial. O cálculo dos açúcares redutores na cana-de-açúcar é feito pela fórmula:

$$\text{AR\%cana} = \text{AR\%caldo} \times (1 - 0{,}01 \times \text{Fibra}) \times C,$$

onde

AR%caldo = (3,641 – 0,0343 x Pureza). A pureza é a porcentagem de sacarose (Pol) contida nos sólidos solúveis (Brix), sendo o principal indicador de maturação da cana-de-açúcar, obtida pela fórmula: Pureza = [(Pol % cana) / (Brix % cana)] x 100.

• Açúcares totais recuperáveis (ATR) representam a quantidade de açúcares (na forma de açúcares invertidos ou ART) que são recuperados na usina (kg/ton.cana), admitindo-se perdas na lavagem de cana, extração (perda de pol no bagaço), torta de filtros e outras perdas indeterminadas. O ATR é obtido por:

$$ATR = 9{,}5263 \times Pol\%cana + 9{,}05 \times AR.$$

• pH, índice que indica o grau de acidez ou alcalinidade do caldo, cujos valores são obtidos em laboratório.

• Cu (cobre), Al (alumínio), Mg (magnésio), Ca (cálcio) e K (potássio) são compostos inorgânicos encontrados no caldo da cana, cujas análises são realizadas em laboratório.

• Compostos fenólicos e Ácido aconítico são precursores da cor, oriundos da própria planta, têm importância fundamental na qualidade dos produtos, influenciando e afetando a cor do açúcar. As determinações desses compostos nas amostras de caldo de cana são realizadas em laboratório.

• Cor ICUMSA é um dos caracteres empregados na avaliação da qualidade e comercialização do açúcar para exportação. Quanto mais baixa a unidade ICUMSA (U.I), mais claro, ou mais branco, é o açúcar (Simioni et al*.,* 2006). O termo ICUMSA é a sigla da International Commission for Uniform Methods of Sugar Analysis (Comissão Internacional para Métodos Uniformes de Análise de Açúcar). Para o cálculo dos valores em Unidades ICUMSA (U.I.), é utilizada a seguinte fórmula:

$$Cor(U.I.) = \frac{-\log T}{b \times c} \times 1000,$$

em que

log T = Logaritmo da transmitância;

b = Comprimento interno da cubeta (cm);

c = Concentração de sacarose em função do Brix a 20°C (g/mL).

## 2.2. Avaliação da multicolinearidade

Para avaliar possíveis ocorrências de multicolinearidade na matriz de trabalho, foram utilizados os seguintes procedimentos: análise da matriz de correlação, fatores de inflação da variância, análise dos autovalores e autovetores da matriz, decomposição em valores singulares e determinante de $X'X$.

A análise da matriz de correlação é um procedimento que envolve a análise dos elementos não-diagonais ($r_{ij}$) da matriz de trabalho. Se as variáveis independentes ($x_i, x_j$) apresentarem dependência linear aproximada entre si, então, $|r_{ij}|$ será próximo de 1. Contudo, a ausência de correlação alta entre quaisquer pares de variáveis considerados não evidencia ausência de multicolinearidade, principalmente quando está envolvido grande número de variáveis (Cruz e Carneiro, 2006).

Os elementos diagonais $C_{jj}$ da matriz $C = (X'X)^{-1}$, denominados por Marquardt (1970) de fatores de inflação da variância (VIFs), constituem um importante diagnóstico da multicolinearidade, apresentado em Montgomery e Peck (1992). A ocorrência de qualquer VIF com valor superior a 10 constitui indicativo de possíveis efeitos adversos provocados pela multicolinearidade, sobre os estimadores em uso, de acordo com Neter et al. (1990).

A análise dos autovalores e autovetores da matriz de correlação é um procedimento eficaz de diagnóstico de multicolinearidade. Quando existe uma ou mais dependências lineares entre as variáveis, um ou mais autovalores ($\lambda_1, \lambda_2, ..., \lambda_p$) da matriz de correlação serão pequenos, segundo Belsley et al. (1980). Se um autovalor é próximo de zero, indicando uma dependência linear entre as observações, os elementos do autovetor associado a esse autovalor descrevem a natureza dessa dependência linear (Carvalho e Cruz, 1996).

Montgomery e Peck (1992) notam que o exame do número de condição (*NC*) de $X'X$**,** definido como a razão entre o maior e o menor autovalor dessa matriz é um critério para a classificação da

multicolinearidade. Esses autores argumentam que, em geral, se o número de condição é menor que 100, não existe problema sério com a multicolinearidade; para um valor de *NC* entre 100 e 1000 implica em multicolinearidade moderada a forte; se o número de condição for maior que 1000, constitui indício de multicolinearidade severa.

Outro procedimento de diagnóstico é a decomposição em valores singulares. De acordo com Lawson e Hanson (1974) e Belsley et al. (1980), qualquer matriz *X nxp*, sendo *n* observações e *p* variáveis, pode ser decomposta da seguinte forma $X = UDT'$, onde $U'U = T'T = I_p$ e *D* é uma matriz diagonal *pxp* com elementos diagonais não-negativos $\mu_j$, $j = 1,2,...,p$, denominados valores singulares de *X*. Então, $X = UDT'$ é a forma de decomposição de *X* em seus valores singulares. Montgomery e Peck (1992) notam que a matriz *X* mal-condicionada afeta o tamanho dos valores singulares, havendo um valor singular pequeno para cada dependência linear aproximada. A magnitude do mal-condicionamento depende de quão pequeno é o valor singular mínimo em relação ao valor singular máximo.

A avaliação do determinante de $X'X$ é também utilizada para diagnosticar a multicolinearidade. Quando a matriz $X'X$ está na forma de correlação, a possível variação dos valores do determinante é $0 \le |X'X| \le 1$. Se $|X'X| = 1$, as variáveis independentes são ortogonais. Se $|X'X| = 0$, existe uma dependência linear completa entre essas variáveis, e à medida que $|X'X|$ se aproxima de zero, a multicolinearidade se torna mais intensa (Montgomery e Peck,1992).

## 2.3. Estimação dos coeficientes de trilha quando ocorre a multicolinearidade

Para atenuar os efeitos adversos da multicolinearidade, na estimação dos coeficientes de trilha, foram utilizadas duas metodologias: o método denominado “ridge regression” (regressão em crista ou em cumeeira,

conforme Cruz e Carneiro, 2006), originalmente proposto por Hoerl e Kennard (1970a,b); e a regressão em componentes principais, segundo Montgomery e Peck (1992).

Na regressão em crista os coeficientes de trilha foram obtidos pela solução da equação

$$(X'X + kI)\theta^* = X'Y$$

em que $X'X$ é a matriz de correlações entre as variáveis independentes do modelo de regressão; $k$ é uma constante, em geral, pertencente ao intervalo $[0,1]$, adicionada aos elementos da diagonal da matriz $X'X$; $I$ é a matriz identidade; $\theta^*$ é o vetor dos estimadores dos coeficientes de trilha; e $X'Y$ é a matriz de correlações entre a variável dependente com cada variável independente do modelo de regressão. O valor adequado referente à constante $k$ foi determinado, neste ensaio, pelo exame do traço de crista (Hoerl e Kennard, 1970b). O traço de crista foi obtido plotando os parâmetros estimados (coeficientes de trilha) em função dos valores de $k$ no intervalo de $[0,1]$. O menor valor de $k$ capaz de estabilizar a maioria dos estimadores dos coeficientes de trilha foi empregado.

Na exclusão de variáveis, por componentes principais, aqueles componentes principais correspondentes aos autovalores próximos de zero foram removidos e a análise de trilha foi aplicada às variáveis restantes.

Assim, para a exclusão de variáveis, foi realizada a análise com base em componentes principais e identificada a variável que apresentou maior valor no autovetor associado ao componente principal de menor autovalor, sendo essa variável excluída. A seguir, foi feito o diagnóstico de multicolinearidade a fim de verificar o seu grau na nova matriz. Sendo constatados, ainda, os efeitos da multicolinearidade, foi realizada uma nova análise, dessa forma, foi, novamente, identificada e excluída a variável que apresentou maior valor no autovetor associado ao componente principal de menor autovalor. Esse processo foi repetido até obter um grau de multicolinearidade considerado fraco, que não constitui problemas sérios. O grau de multicolinearidade da matriz $X'X$ foi estabelecido de acordo com os critérios indicados por Montgomery e Peck (1992), que se baseiam no valor

do número de condição (*NC* = razão entre o maior e o menor autovalor de $X'X$).

O método baseado na regressão em crista e a exclusão de variáveis foram comparados por meio dos coeficientes de trilha obtidos para os caracteres avaliados. Todas as análises foram efetuadas com o auxílio do programa computacional Genes (Cruz, 2009).

## 3. RESULTADOS E DISCUSSÃO

As estimativas dos coeficientes de correlação entre Cor ICUMSA (variável principal) e as variáveis explicativas Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos, Ácido aconítico, foram significativas a 1% e 5% pelo teste t (Tabela 1), sendo essas estimativas positivas e negativas, e a maioria relativamente elevadas.

Pelo exame da matriz de correlações, constata-se que a maior correlação ocorreu entre as variáveis Pol x ATR (0,9984). Esse valor se aproxima da unidade, portanto, é possível que esteja ocorrendo alto grau de colinearidade entre as variáveis. Contudo, os demais métodos devem ser aplicados para que essa hipótese seja confirmada.

**Tabela 1.** Estimativas dos coeficientes de correlação entre caracteres Cor ICUMSA, Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos, Ácido aconítico avaliados no caldo de cana-de-açúcar.

| Caracteres | Brix | Pol | AR | ATR | pH | Cu | Al | Mg | Ca | K | Compostos fenólicos | Ácido aconítico |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cor ICUMSA | -0,8450** | -0,8341** | 0,6757** | -0,8433** | -0,8202** | 0,8556** | 0,8049** | 0,7319** | 0,5440* | 0,8312** | 0,5780* | 0,5074* |
| Brix | | 0,9918** | -0,8463** | 0,9970** | 0,8825** | -0,9360** | -0,8995** | -0,8454** | -0,4641 | -0,8502** | -0,5416* | -0,5637* |
| Pol | | | -0,9047** | 0,9984** | 0,8634** | -0,9058** | -0,8694** | -0,8070** | -0,4400 | -0,8041** | -0,5841* | -0,5632* |
| AR | | | | -0,8791** | -0,6719** | 0,6789** | 0,6818** | 0,5666* | 0,2454 | 0,5248* | 0,6556** | 0,5113* |
| ATR | | | | | 0,8766** | -0,9231** | -0,8820** | -0,8275** | -0,4596 | -0,8298** | -0,5663* | -0,5620* |
| pH | | | | | | -0,9361** | -0,7861** | -0,8768** | -0,7024** | -0,9089** | -0,5088* | -0,5235* |
| Cu | | | | | | | 0,8619** | 0,9239** | 0,6410** | 0,9575** | 0,5099* | 0,5281* |
| Al | | | | | | | | 0,7671** | 0,3245 | 0,7838** | 0,4732 | 0,5501* |
| Mg | | | | | | | | | 0,5158* | 0,8335** | 0,4245 | 0,4522 |
| Ca | | | | | | | | | | 0,7693** | 0,2691 | 0,4071 |
| K | | | | | | | | | | | 0,4349 | 0,4992* |
| Compostos fenólicos | | | | | | | | | | | | 0,6497** |

**,* Significativo a 1% e a 5% pelo teste t, respectivamente.

A Tabela 2 apresenta os resultados dos testes para diagnóstico de multicolinearidade aplicados à matriz de correlações correspondente às variáveis explicativas Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos, Ácido aconítico.

**Tabela 2.** Diagnóstico de multicolinearidade da matriz de correlação $(X'X)$ envolvendo as variáveis explicativas Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos, Ácido aconítico.

| Ordem | Autovalores | Valor singular | Índice de condição | VIF |
|---|---|---|---|---|
| 1 | 8,8302 | 2,9716 | 1 | 775,447 |
| 2 | 1,2037 | 1,0971 | 2,7085 | - |
| 3 | 0,8974 | 0,9473 | 3,1369 | - |
| 4 | 0,4246 | 0,6516 | 4,5604 | - |
| 5 | 0,3163 | 0,5624 | 5,2839 | 15,8185 |
| 6 | 0,1873 | 0,4327 | 6,8670 | 254,1085 |
| 7 | 0,0826 | 0,2873 | 10,3417 | 12,7769 |
| 8 | 0,0400 | 0,2001 | 14,8532 | 48,1777 |
| 9 | 0,0155 | 0,1245 | 23,8701 | 12,6058 |
| 10 | 0,0017 | 0,0415 | 71,5836 | 216,5727 |
| 11 | 0,0008 | 0,0283 | 104,8491 | 4,7194 |
| 12 | 0,0000 | 0,0027 | 1102,355 | 3,1433 |

Determinante: 0,0 Número de condição (*NC*): 1215187,160574

O número de condição (*NC*) encontrado foi excessivamente alto, assim, a multicolinearidade pode ser classificada como severa, conforme critério de Montgomery e Peck (1992). Verifica-se que alguns autovalores apresentam estimativas iguais a zero ou próximas de zero, sugerindo a existência de diversas relações lineares determinantes de efeitos prejudiciais de multicolinearidade.

A decomposição da matriz de correlações em valores singulares e a determinação dos índices de condição, dados pela relação entre o maior e os demais valores singulares, fornecem também informações sobre o grau de multicolinearidade presente. Valores singulares pequenos e índice de condição elevado indicam problemas sérios proporcionados pela

multicolinearidade. Pela Tabela 2, verifica-se que o valor singular e o índice de condição, associados à 10ª, 11ª e 12ª ordens, apresentam estimativas que traduzem a existência de problemas de multicolinearidade.

Foram encontrados 10 valores dos fatores de inflação da variância (VIFs) superiores a 10, em valor absoluto, indicando que está ocorrendo multicolinearidade em grau elevado.

Nas matrizes de correlação, o determinante varia de zero a um, caso as variáveis sejam perfeitamente correlacionadas ou ortogonais entre si, respectivamente. O determinante correspondente à matriz em estudo foi nulo, consistindo num indicativo da existência de problemas de multicolinearidade nessa matriz. O próprio diagnóstico por certos critérios se torna problemático, pois alguns deles, como VIF, dependem de elementos da matriz inversa que pode ser gerada, apesar de singular, por meio de um processamento computacional com erro numérico.

As causas da multicolinearidade podem ser avaliadas com base nas informações apresentadas na Tabela 3. A análise dos elementos dos autovetores associados aos últimos autovalores, de menor magnitude, indica que Pol, Brix e Cu são características que, em combinações com outras, podem estar proporcionando os problemas na matriz estudada.

**Tabela 3.** Autovalores e autovetores da matriz de correlação $(X'X)$ entre as variáveis explicativas Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos, Ácido aconítico.

| Ordem | Autovalores | Autovetores | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Brix | Pol | AR | ATR | pH | Cu | Al | Mg | Ca | K | Compostos fenólicos | Ácido aconítico |
| 1 | 8,8302 | 0,3279 | 0,3245 | -0,2709 | 0,327 | 0,3161 | -0,3262 | -0,2975 | -0,2957 | -0,2007 | -0,3048 | -0,2108 | -0,2166 |
| 2 | 1,2037 | 0,07 | 0,1417 | -0,4164 | 0,1031 | -0,2014 | 0,1713 | -0,0834 | 0,1883 | 0,5533 | 0,3348 | -0,4442 | -0,2512 |
| 3 | 0,8974 | 0,1886 | 0,1734 | -0,1248 | 0,1774 | -0,017 | -0,0713 | -0,2285 | -0,1508 | 0,4636 | 0,0535 | 0,4823 | 0,5972 |
| 4 | 0,4246 | 0,0062 | -0,1201 | 0,4492 | -0,0746 | -0,1108 | -0,0494 | -0,4543 | -0,194 | 0,3441 | -0,044 | 0,2755 | -0,5686 |
| 5 | 0,3163 | -0,1315 | -0,1765 | 0,3207 | -0,1545 | 0,0812 | -0,1549 | -0,0221 | -0,4599 | 0,2732 | -0,0698 | -0,6139 | 0,3549 |
| 6 | 0,1873 | -0,0087 | 0,0319 | -0,1531 | 0,0151 | 0,134 | 0,0317 | 0,5498 | -0,6426 | 0,0699 | 0,3478 | 0,2248 | -0,2544 |
| 7 | 0,0826 | -0,1959 | -0,1377 | -0,1026 | -0,1672 | 0,7682 | 0,2398 | -0,3555 | 0,0709 | -0,0306 | 0,3499 | 0,0335 | 0,0278 |
| 8 | 0,0400 | 0,1899 | 0,0796 | 0,2645 | 0,1239 | 0,4735 | 0,0714 | 0,4542 | 0,3194 | 0,4311 | -0,373 | 0,0102 | -0,0977 |
| 9 | 0,0155 | 0,3671 | 0,2293 | 0,4356 | 0,3132 | -0,0592 | 0,6134 | -0,0677 | -0,1178 | -0,2208 | 0,2242 | -0,1373 | 0,0936 |
| 10 | 0,0017 | 0,1034 | 0,1147 | 0,3456 | 0,1774 | 0,0678 | -0,6248 | 0,0756 | 0,2664 | -0,0643 | 0,5893 | -0,0446 | 0,0103 |
| 11 | 0,0008 | 0,784 | -0,4008 | -0,1109 | -0,4499 | 0,0019 | -0,0491 | 0,0157 | 0,0207 | -0,0333 | 0,0723 | -0,0147 | 0,016 |
| 12 | 0,0000 | -0,0067 | -0,7391 | -0,0877 | 0,6678 | -0,0003 | 0,0022 | -0,0001 | -0,0008 | 0,0003 | -0,0018 | 0,0002 | -0,0002 |

Tendo em vista os resultados obtidos no diagnóstico de multicolinearidade, foram realizadas duas análises, uma incluindo todas as variáveis, com as estimativas obtidas pelo método baseado em regressão em crista; e a outra com exclusão de variáveis, baseada na regressão em componentes principais.

O valor de $k$ para as análises de trilha com base no método de regressão em crista foi determinado pelo exame do traço de crista apresentado na Figura 1. Foi admitido que as estimativas estavam estabilizadas a partir de um valor de $k$ igual a 0,1. De acordo com Cruz e Carneiro (2006), com o método de crista é esperado que as estimativas dos efeitos diretos e indiretos sejam tendenciosas, porém associadas a menores VIFs.

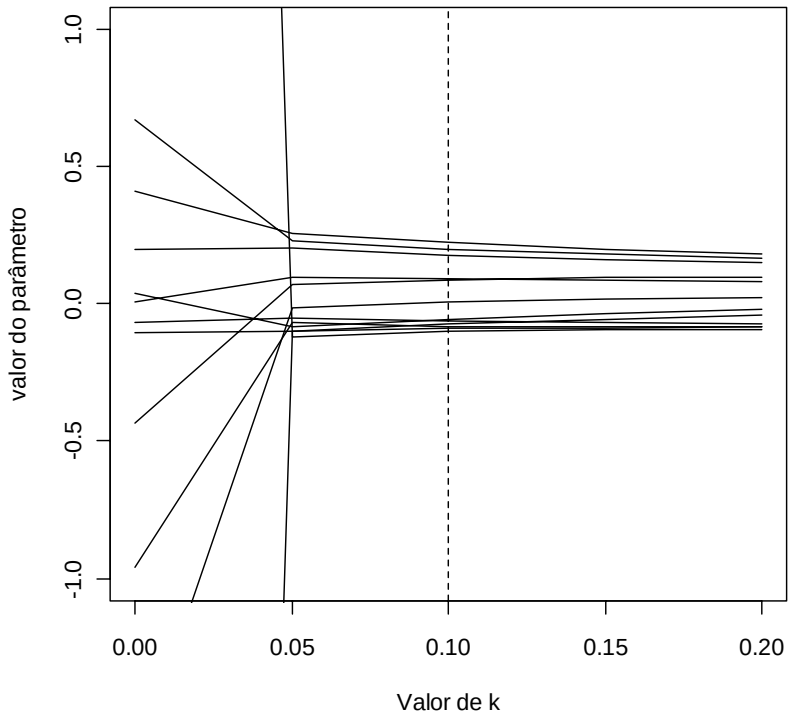


**Figura 1.** Gráfico de crista para as estimativas dos coeficientes de trilha entre as variáveis explicativas Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos, Ácido aconítico e a variável básica Cor ICUMSA, correspondentes aos valores de $k$ no intervalo $0 < k < 1$. O valor $k = 0{,}1$ foi usado para a estimação dos coeficientes de trilha.

O estimador de componentes principais reduz os efeitos da multicolinearidade por usar um subconjunto de componentes principais no modelo (Montgomery e Peck, 1992). Assim, foram identificadas, nesta ordem, as variáveis Pol, ATR, Cu, Brix e K, como aquelas que podem ser descartadas do estudo realizado para contornar os efeitos da multicolinearidade. No entanto, estas características estão sendo apontadas por um critério estatístico.

Carvalho et al. (1999), em estudo de análise de trilha sob multicolinearidade em pimentão, mencionam que no descarte de variáveis não necessariamente a que tem maior importância econômica é a que mais explica a variável básica, o que torna difícil o processo de exclusão, principalmente quando várias variáveis necessitam ser eliminadas.

Após a exclusão dessas variáveis, foi realizado o diagnóstico de multicolinearidade conforme critérios citados anteriormente, resultando em multicolinearidade fraca (*NC*: Número de condição = 84,25).

Os resultados das análises de trilha aplicando os dois métodos estão apresentados na Tabela 4.

**Tabela 4.** Estimativas dos efeitos diretos e indiretos das variáveis explicativas Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos, Ácido aconítico sobre a variável básica Cor ICUMSA, obtidas pelo método baseado na regressão em crista (Método 1) e com exclusão de variáveis por componentes principais (Método 2).

| | Método 1 | | Método 2 | |
|---|---|---|---|---|
| | k = 0,1 | | k = 0,0 | |
| Variáveis | Efeito | VIF | Efeito | VIF |
| Brix | | | | |
| Efeito direto sobre Cor ICUMSA | -0,0816 | 9,5811 | | |
| Efeito indireto via Pol | -0,0890 | 7,8072 | | |
| Efeito indireto via AR | -0,0024 | 3,1792 | | |
| Efeito indireto via ATR | -0,1009 | 7,9488 | | |
| Efeito indireto via pH | -0,0579 | 3,9298 | | |
| Efeito indireto via Cu | -0,0871 | 6,6211 | | |
| Efeito indireto via Al | -0,2026 | 3,0537 | | |
| Efeito indireto via Mg | 0,0490 | 2,6099 | | |
| Efeito indireto via Ca | -0,0432 | 0,5773 | | |
| Efeito indireto via K | -0,1648 | 4,4016 | | |
| Efeito indireto via Compostos Fenólicos | -0,0972 | 0,5298 | | |
| Efeito indireto via Ácido Aconítico | 0,0409 | 0,5422 | | |
| Total | -0,845 | | | |
| Pol | | | | |
| Efeito direto sobre Cor ICUMSA | -0,0898 | 9,6071 | | |
| Efeito indireto via Brix | -0,0810 | 7,7860 | | |
| Efeito indireto via AR | -0,0026 | 3,6334 | | |
| Efeito indireto via ATR | -0,1010 | 7,9712 | | |
| Efeito indireto via pH | -0,0567 | 3,7618 | | |
| Efeito indireto via Cu | -0,0843 | 6,2009 | | |
| Efeito indireto via Al | -0,1958 | 2,8529 | | |
| Efeito indireto via Mg | 0,0468 | 2,3780 | | |
| Efeito indireto via Ca | -0,0410 | 0,5189 | | |
| Efeito indireto via K | -0,1559 | 3,9375 | | |
| Efeito indireto via Compostos Fenólicos | -0,1048 | 0,6163 | | |
| Efeito indireto via Ácido Aconítico | 0,0409 | 0,5412 | | |
| Total | -0,8341 | | | |
| AR | | | | |
| Efeito direto sobre Cor ICUMSA | 0,0028 | 5,3734 | 0,104 | 3,1087 |
| Efeito indireto via Brix | 0,0691 | 5,6688 | | |
| Efeito indireto via Pol | 0,0812 | 6,4961 | | |
| Efeito indireto via ATR | 0,0889 | 6,1799 | | |
| Efeito indireto via pH | 0,0441 | 2,2778 | 0,0322 | 6,0771 |
| Efeito indireto via Cu | 0,0632 | 3,4832 | | |
| Efeito indireto via Al | 0,1536 | 1,7546 | 0,3844 | 1,9985 |
| Efeito indireto via Mg | -0,0329 | 1,1722 | 0,0111 | 1,6453 |
| Efeito indireto via Ca | 0,0229 | 0,1614 | 0,0724 | 0,2170 |
| Efeito indireto via K | 0,1017 | 1,6771 | | |
| Efeito indireto via Compostos Fenólicos | 0,1176 | 0,7764 | 0,1555 | 1,0162 |
| Efeito indireto via Ácido Aconítico | -0,0371 | 0,4460 | -0,0839 | 0,5888 |
| Total | 0,6757 | | 0,6757 | |

Continuação...

Tabela, continuação.

| | Método 1 | | Método 2 | |
|---|---|---|---|---|
| | k = 0,1 | | k = 0,0 | |
| Variáveis | Efeito | VIF | Efeito | VIF |
| ATR | | | | |
| Efeito direto sobre Cor ICUMSA | -0,1012 | 9,6800 | | |
| Efeito indireto via Brix | -0,0814 | 7,8676 | | |
| Efeito indireto via Pol | -0,0896 | 7,9112 | | |
| Efeito indireto via AR | -0,0025 | 3,4305 | | |
| Efeito indireto via pH | -0,0575 | 3,8774 | | |
| Efeito indireto via Cu | -0,0859 | 6,4397 | | |
| Efeito indireto via Al | -0,1987 | 2,9359 | | |
| Efeito indireto via Mg | 0,0480 | 2,5007 | | |
| Efeito indireto via Ca | -0,0428 | 0,5661 | | |
| Efeito indireto via K | -0,1609 | 4,1933 | | |
| Efeito indireto via Compostos Fenólicos | -0,1016 | 0,5793 | | |
| Efeito indireto via Ácido Aconítico | 0,0408 | 0,5388 | | |
| Total | -0,8433 | | | |
| pH | | | | |
| Efeito direto sobre Cor ICUMSA | -0,0656 | 6,1081 | -0,0479 | 13,4629 |
| Efeito indireto via Brix | -0,0720 | 6,1642 | | |
| Efeito indireto via Pol | -0,0775 | 5,9167 | | |
| Efeito indireto via AR | -0,0019 | 2,0038 | -0,0699 | 1,4032 |
| Efeito indireto via ATR | -0,0887 | 6,1447 | | |
| Efeito indireto via Cu | -0,0871 | 6,6226 | | |
| Efeito indireto via Al | -0,1771 | 2,3324 | -0,4432 | 2,6566 |
| Efeito indireto via Mg | 0,0508 | 2,8075 | -0,0172 | 3,9406 |
| Efeito indireto via Ca | -0,0654 | 1,3225 | -0,2073 | 1,7788 |
| Efeito indireto via K | -0,1762 | 5,0306 | | |
| Efeito indireto via Compostos Fenólicos | -0,0913 | 0,4677 | -0,1207 | 0,6121 |
| Efeito indireto via Ácido Aconítico | 0,0380 | 0,4677 | 0,0859 | 0,6175 |
| Total | -0,8202 | | -0,8202 | |
| Cu | | | | |
| Efeito direto sobre Cor ICUMSA | 0,0931 | 9,1482 | | |
| Efeito indireto via Brix | 0,0764 | 6,9344 | | |
| Efeito indireto via Pol | 0,0813 | 6,5120 | | |
| Efeito indireto via AR | 0,0019 | 2,0460 | | |
| Efeito indireto via ATR | 0,0934 | 6,8140 | | |
| Efeito indireto via pH | 0,0614 | 4,4218 | | |
| Efeito indireto via Al | 0,1941 | 2,8038 | | |
| Efeito indireto via Mg | -0,0536 | 3,1172 | | |
| Efeito indireto via Ca | 0,0597 | 1,1013 | | |
| Efeito indireto via K | 0,1856 | 5,5829 | | |
| Efeito indireto via Compostos Fenólicos | 0,0915 | 0,4697 | | |
| Efeito indireto via Ácido Aconítico | -0,0384 | 0,4758 | | |
| Total | 0,8556 | | | |

Continuação...

Tabela, continuação.

| | Método 1 | | Método 2 | |
|---|---|---|---|---|
| | k = 0,1 | | k = 0,0 | |
| Variáveis | Efeito | VIF | Efeito | VIF |
| Al | | | | |
| Efeito direto sobre Cor ICUMSA | 0,2253 | 4,5686 | 0,5638 | 4,2987 |
| Efeito indireto via Brix | 0,0734 | 6,4041 | | |
| Efeito indireto via Pol | 0,0780 | 5,9993 | | |
| Efeito indireto via AR | 0,0019 | 2,0637 | 0,0709 | 1,4452 |
| Efeito indireto via ATR | 0,0892 | 6,2207 | | |
| Efeito indireto via pH | 0,0516 | 3,1184 | 0,0377 | 8,3201 |
| Efeito indireto via Cu | 0,0802 | 5,6144 | | |
| Efeito indireto via Mg | -0,0445 | 2,1487 | 0,0150 | 3,0159 |
| Efeito indireto via Ca | 0,0302 | 0,2822 | 0,0958 | 0,3796 |
| Efeito indireto via K | 0,1520 | 3,7408 | | |
| Efeito indireto via Compostos Fenólicos | 0,0849 | 0,4044 | 0,1122 | 0,5294 |
| Efeito indireto via Ácido Aconítico | -0,0400 | 0,5163 | -0,0903 | 0,6817 |
| Total | 0,8049 | | 0,8049 | |
| Mg | | | | |
| Efeito direto sobre Cor ICUMSA | -0,0580 | 4,4204 | 0,0196 | 5,1257 |
| Efeito indireto via Brix | 0,0690 | 5,6569 | | |
| Efeito indireto via Pol | 0,0724 | 5,1683 | | |
| Efeito indireto via AR | 0,0016 | 1,4249 | 0,0589 | 0,9979 |
| Efeito indireto via ATR | 0,0837 | 5,4762 | | |
| Efeito indireto via pH | 0,0575 | 3,8794 | 0,0420 | 10,3503 |
| Efeito indireto via Cu | 0,0860 | 6,4510 | | |
| Efeito indireto via Al | 0,1728 | 2,2207 | 0,4325 | 2,5294 |
| Efeito indireto via Ca | 0,0480 | 0,7130 | 0,1522 | 0,9590 |
| Efeito indireto via K | 0,1616 | 4,2304 | | |
| Efeito indireto via Compostos Fenólicos | 0,0762 | 0,3255 | 0,1007 | 0,4261 |
| Efeito indireto via Ácido Aconítico | -0,0328 | 0,3489 | -0,0742 | 0,4607 |
| Total | 0,7319 | | 0,7319 | |
| Ca | | | | |
| Efeito direto sobre Cor ICUMSA | 0,0931 | 3,2444 | 0,2951 | 3,6051 |
| Efeito indireto via Brix | 0,0379 | 1,7049 | | |
| Efeito indireto via Pol | 0,0395 | 1,5365 | | |
| Efeito indireto via AR | 0,0007 | 0,2672 | 0,0255 | 0,1871 |
| Efeito indireto via ATR | 0,0465 | 1,6889 | | |
| Efeito indireto via pH | 0,0461 | 2,4897 | 0,0337 | 6,6426 |
| Efeito indireto via Cu | 0,0596 | 3,1054 | | |
| Efeito indireto via Al | 0,0731 | 0,3974 | 0,1829 | 0,4527 |
| Efeito indireto via Mg | -0,0299 | 0,9714 | 0,0101 | 1,3634 |
| Efeito indireto via K | 0,1491 | 3,6038 | | |
| Efeito indireto via Compostos Fenólicos | 0,0483 | 0,1308 | 0,0638 | 0,1712 |
| Efeito indireto via Ácido Aconítico | -0,0296 | 0,2827 | -0,0668 | 0,3733 |
| Total | 0,544 | | 0,544 | |

Continuação...

Tabela, continuação.

| | Método 1 | | Método 2 | |
|---|---|---|---|---|
| | k = 0,1 | | k = 0,0 | |
| Variáveis | Efeito | VIF | Efeito | VIF |
| K | | | | |
| Efeito direto sobre Cor ICUMSA | 0,1939 | 7,3714 | | |
| Efeito indireto via Brix | 0,0694 | 5,7210 | | |
| Efeito indireto via Pol | 0,0722 | 5,1318 | | |
| Efeito indireto via AR | 0,0015 | 1,2225 | | |
| Efeito indireto via ATR | 0,0839 | 5,5066 | | |
| Efeito indireto via pH | 0,0597 | 4,1685 | | |
| Efeito indireto via Cu | 0,0891 | 6,9287 | | |
| Efeito indireto via Al | 0,1766 | 2,3185 | | |
| Efeito indireto via Mg | -0,0483 | 2,5369 | | |
| Efeito indireto via Ca | 0,0716 | 1,5862 | | |
| Efeito indireto via Compostos Fenólicos | 0,0780 | 0,3416 | | |
| Efeito indireto via Ácido Aconítico | -0,0363 | 0,4252 | | |
| Total | 0,8312 | | | |
| Compostos Fenólicos | | | | |
| Efeito direto sobre Cor ICUMSA | 0,1794 | 2,1866 | 0,2372 | 2,3644 |
| Efeito indireto via Brix | 0,0442 | 2,3215 | | |
| Efeito indireto via Pol | 0,0524 | 2,7080 | | |
| Efeito indireto via AR | 0,0019 | 1,9079 | 0,0682 | 1,3361 |
| Efeito indireto via ATR | 0,0573 | 2,5646 | | |
| Efeito indireto via pH | 0,0334 | 1,3064 | 0,0244 | 3,4855 |
| Efeito indireto via Cu | 0,0474 | 1,9649 | | |
| Efeito indireto via Al | 0,1066 | 0,8450 | 0,2668 | 0,9625 |
| Efeito indireto via Mg | -0,0246 | 0,6581 | 0,0083 | 0,9236 |
| Efeito indireto via Ca | 0,0251 | 0,1941 | 0,0794 | 0,2610 |
| Efeito indireto via K | 0,0843 | 1,1515 | | |
| Efeito indireto via Ácido Aconítico | -0,0472 | 0,7202 | -0,1066 | 0,9509 |
| Total | 0,578 | | 0,578 | |
| Ácido Aconítico | | | | |
| Efeito direto sobre Cor ICUMSA | -0,0726 | 2,0653 | -0,1641 | 2,2527 |
| Efeito indireto via Brix | 0,0460 | 2,5154 | | |
| Efeito indireto via Pol | 0,0505 | 2,5175 | | |
| Efeito indireto via AR | 0,0015 | 1,1604 | 0,0532 | 0,8126 |
| Efeito indireto via ATR | 0,0569 | 2,5254 | | |
| Efeito indireto via pH | 0,0344 | 1,3831 | 0,0251 | 3,6902 |
| Efeito indireto via Cu | 0,0491 | 2,1078 | | |
| Efeito indireto via Al | 0,1239 | 1,1421 | 0,3101 | 1,3008 |
| Efeito indireto via Mg | -0,0262 | 0,7468 | 0,0089 | 1,0482 |
| Efeito indireto via Ca | 0,0379 | 0,4441 | 0,1201 | 0,5974 |
| Efeito indireto via K | 0,0968 | 1,5175 | | |
| Efeito indireto via Compostos Fenólicos | 0,1166 | 0,7625 | 0,1541 | 0,9980 |
| Total | 0,5074 | | 0,5074 | |
| Coeficiente de determinação ($R^2$) | | 0,7820 | | 0,7921 |
| Efeito da variável residual | | 0,4669 | | 0,4560 |
| Determinante da matriz $X'X$ | | $2,23x10^{-6}$ | | $3,0x10^{-3}$ |

A análise de trilha, após a exclusão de variáveis por meio da regressão em componentes principais, mostrou resultados semelhantes à análise de trilha baseada na regressão em crista. Os fatores de inflação das variâncias (VIFs) foram pequenos, para ambos os métodos, mostrando serem confiáveis em expressar as verdadeiras relações de causa e efeito entre os caracteres estudados.

O coeficiente de determinação obtido na análise de trilha baseada no método de regressão em crista (0,7820) foi semelhante ao encontrado pelo método da regressão em componentes principais (0,7921).

Nos dois métodos, os efeitos diretos e indiretos das variáveis explicativas sobre Cor ICUMSA oscilaram entre valores positivos e negativos, e não foram superiores à unidade.

Na regressão em crista e na regressão em componentes principais, os efeitos diretos da variável Al (0,225 e 0,5638, respectivamente) foram superiores aos indiretos. As variáveis K e Compostos fenólicos apresentaram situação semelhante na regressão em crista, com efeitos diretos iguais a 0,1939 e 0,1794, respectivamente. A variável Compostos fenólicos mostrou efeito direto igual a 0,2372 na análise baseada em componentes principais, porém, teve um efeito indireto via Al maior (0,2668).

Para efeito de seleção, é importante identificar, dentre os caracteres de alta correlação com a variável principal, aqueles de maior efeito direto em sentido favorável à seleção, de tal forma que a resposta correlacionada por meio da seleção indireta seja eficiente (Cruz et al., 2004). Portanto, Al, K e Compostos fenólicos são as características que mais se associam à Cor. Assim, quanto maior o teor de Al, K e/ou Compostos fenólicos, mais elevada será a Cor do caldo.

Segundo Santos (2008), o excesso de K não é desejável para produção de açúcar, pois como é o maior constituinte das cinzas (substâncias inorgânicas), está em altas concentrações no caldo, dificultando a cristalização, reduzindo o rendimento industrial. Em um estudo para comprovar a correlação da quantidade de Compostos fenólicos totais

com a cor do açúcar produzido, Simioni et al., (2006) relatam que quanto maior a concentração de Compostos fenólicos, maior será a Cor ICUMSA.

Observa-se também que o Al é a variável que apresenta maior influência indireta sobre as demais variáveis explicativas, seguida de K.

As variáveis AR e Cu apresentaram baixo efeito direto e correlação relativamente alta, de mesmo sinal em relação ao Al. De acordo com Santos (2008), embora AR não seja o principal determinante que explique a Cor do caldo, é uma característica diretamente relacionada com essa variável principal. Zarpelon (1988), citado por Santos (2008), salienta ser conhecido o fato que os açúcares redutores do caldo elevam a cor do açúcar.

De acordo com Cruz et al. (2004), quando uma variável explicativa apresenta correlação favorável com a variável principal e efeito direto em sentido desfavorável, é indicativo da ausência de causa e efeito, ou seja, o caráter independente não é o principal determinante das alterações na variável principal, existindo outros fatores que poderão proporcionar maior impacto em termos de ganhos de seleção. Tal situação é observada para os caracteres Mg (considerando apenas o resultado encontrado pelo método de regressão em crista) e Ácido aconítico que apresentaram correlação positiva com Cor do caldo e efeito direto em sentido contrário. Conforme Stupiello (2001), citado por Santos (2008), quanto mais altas as concentrações de ácido aconítico no caldo, pior será para clarificação devido à competição com o ácido fosfórico pelo cálcio.

Todas as variáveis avaliadas apresentaram baixos valores de efeitos diretos, sendo estes, inferiores ao valor de efeito da variável residual, com exceção da variável Al (ao considerar o resultado obtido com a exclusão de variáveis).

As variáveis Brix, Pol, ATR e pH mostraram relação inversa com a Cor do caldo, com estimativas de correlação elevadas e efeitos diretos negativos e baixos.

Diante dos resultados encontrados, é plausível sugerir que as variáveis explicativas não devem ser totalmente descartadas devido ao baixo efeito direto sobre a variável principal. Segundo Cruz et al., 2004, caracteres

com alta correlação, mas com baixo efeito direto indicam que a melhor estratégia deverá ser a seleção simultânea de caracteres, com ênfase também nos caracteres cujos efeitos indiretos são significativos.

# 4. CONCLUSÕES

Os procedimentos utilizados para detectar a multicolinearidade mostraram ser eficientes para a quantificação da intensidade com que a multicolinearidade se manifesta e também para a identificação das variáveis envolvidas.

Sob multicolinearidade severa, o método baseado na regressão em crista e a exclusão de variáveis por componentes principais apresentaram resultados semelhantes na estimativa dos coeficientes de trilha, proporcionando sensível redução na magnitude dos fatores de inflação da variância associados aos efeitos diretos e indiretos da análise de trilha.

As variáveis Al, K e Compostos fenólicos são as que melhor explicam a variável Cor do caldo. Contudo, os demais caracteres devem ser levados em consideração devido a elevada correlação existente e a baixa magnitude dos efeitos diretos, evidenciando a necessidade de seleção simultânea de caracteres, com ênfase também nos caracteres cujos efeitos indiretos são significativos.

Todos os caracteres apresentaram alta correlação com a Cor do caldo. No entanto, a análise de trilha revelou que nenhuma correlação representou associação de causa e efeito, tendo em vista as baixas estimativas dos efeitos diretos.

Para fins de melhoramento, a seleção indireta para Cor do caldo, por meio de índice de seleção envolvendo as variáveis Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos e Ácido aconítico é recomendável.

## REFERÊNCIAS BIBLIOGRÁFICAS


BELSLEY, D. A.; KUH, E.; WELSCH, R. E. **Regression Diagnostics: Identifying influential data and sources of collinearity**. New York: John Wiley & Sons, 1980. 292p.

CARVALHO, C. G. P. de; OLIVEIRA, V. R.; CRUZ, C. D.; CASAL, V. W. D. Análise de trilha sob multicolinearidade em pimentão. **Pesquisa Agropecuária Brasileira**., v. 34, n. 4, p. 603-613, 1999.

CARVALHO, S. P. **Métodos alternativos de estimação de coeficientes de trilha e índices de seleção, sob multicolinearidade**. Viçosa, 1995. 163 p. Tese (Doutorado em Genética e Melhoramento). Universidade Federal de Viçosa, Viçosa.

CARVALHO, S. P.; CRUZ, C. D. Diagnosis of multicollinearity: assessment of the condition of correlation matrices used in genetic studies. **Revista Brasileira de Genética**. v. 19, n. 3, p. 479-484, 1996.

CHATTERJEE, S.; PRICE, B. **Regression analysis by example**. New York: John Wiley & Sons, 1977. 228p.

CLARKE, M. A.; LEGENDRE, B. R. Qualidade da cana-de-açúcar: Impactos no rendimento do açúcar e fatores da qualidade. **STAB**, v. 17, n.6, p. 36-40, 1999.

COPERSUCAR. Centro de Tecnologia de Cana (CTC). **Manual de controle químico da fabricação de açúcar**. Piracicaba, 261p, 2001.

CRUZ, C.D. **Programa Genes.** Viçosa: UFV. Versão 2009.

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** v. 2. 2. ed. rev. Viçosa: Editora UFV, 2006.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** v.1. 3. ed. Viçosa: Editora UFV, 2004. 480p.

FERNANDES, A. M.; QUEIROZ, A. C.; PEREIRA, J. C.; LANA, R. P.; BARBOSA, M. H. P.; FONSECA, D. M.; DETMANN, E.; CABRAL, L. S.; PEREIRA, E. S.; VITTOR, A. Composição químico-bromatológica de variedades de cana-de-açúcar (*Saccharum spp* L.*)* com diferentes ciclos de produção (precoce e intermediário) em três idades de corte. **Revista Brasileira de Zootecnia**, v.32, n.4, p.977-985, 2003.

GUNST, R. F.; MASON, R. L. Advantages of examining multicollinearities in regression analysis. **Biometrics**. v. 33, p. 249-260, 1977.

HOERL, A. E. Optimum solution of many variable equations. **Chemical Engineering Progress.** v. 55, p. 69-78, 1959.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**. v. 12, n. 1, p. 55-67, 1970a.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Applications to nonorthogonal problems. **Technometrics**. v. 12, n. 1, p. 69-82, 1970b.

KUTNER, M. H; NACHTSHEIM, C. J.; NETER, J.; LI, W. **Applied linear models**. Boston: Mcgraw-Hill Irwin, 5. ed., 2005. 1396p.

LAWSON, C. L.; HANSON, R. J. **Solving least square problems**. Prentice-Hal: Englewood Cliffs, 1974. 340p.

MARQUARDT, D. W. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. **Technometrics**. v. 13, n.3, p. 591-612, 1970.

MONTGOMERY, D. C.; PECK, E.A. **Introduction to Linear Regression Analysis**. 2. ed., New York: John Wiley & Sons, 1992. 544p.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**. 2. reimpr., Rio de Janeiro: LTC, 2008.

NETER, J.; WASSERMAN, W. **Applied linear statistical models: Regression, analysis of variance, and experimental designs. 3. ed.**. Homewood: Richard D. Irwin, 1990, 1181p.

SANTOS, F. A. **Análise de trilha dos principais constituintes orgânicos e inorgânicos sobre a cor do caldo em cultivares de cana-de-açúcar**. Viçosa, 2008. 50p. Dissertação (Mestrado em Fitotecnia). Universidade Federal de Viçosa, Viçosa.

SIMIONI, K. R.; SILVA, L. F. L. F.; BARBOSA, V.; RÉ, F. E.; BERNARDINO, C. D.; LOPES, M. L.; AMORIM, H. V. Efeito da variedade e época de colheita no teor de fenóis totais em cana-de-açúcar. **STAB**, Piracicaba, v.24, n.3, p.36-39, 2006.

STUPIELLO, J. P. Sarkarana: Importante polissacarídeo. **STAB**, Piracicaba, v.20, n. 1, p.14, 2001.

VITTINGHOFF, E.; GLIDDEN, D. V.; SHIBOSKI, S. C.; MCCULLOCH, C. E. **Regression Methods in Biostatistics: Linear, logistic, survival, and repeated measure models**. New York: Springer. 2005. 340p.

ZARPELON, F. Processamento industrial de cana não despontada: experiência da Usina Estér. **STAB**, v.6, n.6, p.37- 42, 1988.

# APÊNDICE

**Tabela 1.** Estimativas dos efeitos diretos e indiretos das variáveis explicativas Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos, Ácido aconítico sobre a variável básica Cor ICUMSA, obtidas pelo método baseado em análise de regressão em crista (Método 1).

| Método 1 | | |
|---|---|---|
| k = 0,05 | | |
| Variáveis | Efeito | VIF |
| Brix | | |
| Efeito direto sobre Cor ICUMSA | -0,0745 | 16,0307 |
| Efeito indireto via Pol | -0,0975 | 14,5737 |
| Efeito indireto via AR | 0,0136 | 5,1221 |
| Efeito indireto via ATR | -0,1175 | 14,7165 |
| Efeito indireto via pH | -0,0460 | 5,5203 |
| Efeito indireto via Cu | -0,0687 | 11,7054 |
| Efeito indireto via Al | -0,2335 | 4,1304 |
| Efeito indireto via Mg | 0,0710 | 3,6417 |
| Efeito indireto via Ca | -0,0486 | 0,8122 |
| Efeito indireto via K | -0,1848 | 7,4172 |
| Efeito indireto via Compostos Fenólicos | -0,1089 | 0,6232 |
| Efeito indireto via Ácido Aconítico | 0,0543 | 0,6240 |
| Total | -0,845 | |
| Pol | | |
| Efeito direto sobre Cor ICUMSA | -0,0983 | 16,4138 |
| Efeito indireto via Brix | -0,0739 | 14,2336 |
| Efeito indireto via AR | 0,0145 | 5,8537 |
| Efeito indireto via ATR | -0,1177 | 14,7580 |
| Efeito indireto via pH | -0,0450 | 5,2843 |
| Efeito indireto via Cu | -0,0665 | 10,9625 |
| Efeito indireto via Al | -0,2257 | 3,8589 |
| Efeito indireto via Mg | 0,0677 | 3,3182 |
| Efeito indireto via Ca | -0,0461 | 0,7300 |
| Efeito indireto via K | -0,1747 | 6,6353 |
| Efeito indireto via Compostos Fenólicos | -0,1174 | 0,7250 |
| Efeito indireto via Ácido Aconítico | 0,0543 | 0,6228 |
| Total | -0,8341 | |
| AR | | |
| Efeito direto sobre Cor ICUMSA | -0,0160 | 7,9234 |
| Efeito indireto via Brix | 0,0630 | 10,3630 |
| Efeito indireto via Pol | 0,0889 | 12,1263 |
| Efeito indireto via ATR | 0,1036 | 11,4416 |
| Efeito indireto via pH | 0,0350 | 3,1996 |
| Efeito indireto via Cu | 0,0499 | 6,1580 |
| Efeito indireto via Al | 0,1770 | 2,3733 |
| Efeito indireto via Mg | -0,0476 | 1,6357 |
| Efeito indireto via Ca | 0,0257 | 0,2270 |
| Efeito indireto via K | 0,1140 | 2,8261 |
| Efeito indireto via Compostos Fenólicos | 0,1318 | 0,9133 |
| Efeito indireto via Ácido Aconítico | -0,0493 | 0,5132 |
| Total | 0,6757 | |
| ATR | | |
| Efeito direto sobre Cor ICUMSA | -0,1179 | 16,4028 |
| Efeito indireto via Brix | -0,0743 | 14,3826 |

Tabela, continuação.

Método 1
k = 0,05
Variáveis Efeito VIF
Continuação,,,

| | | |
|---|---|---|
| Efeito indireto via Pol | -0,0982 | 14,7678 |
| Efeito indireto via AR | 0,0141 | 5,5268 |
| Efeito indireto via pH | -0,0457 | 5,4466 |
| Efeito indireto via Cu | -0,0678 | 11,3847 |
| Efeito indireto via Al | -0,2289 | 3,9712 |
| Efeito indireto via Mg | 0,0695 | 3,4894 |
| Efeito indireto via Ca | -0,0482 | 0,7963 |
| Efeito indireto via K | -0,1803 | 7,0662 |
| Efeito indireto via Compostos Fenólicos | -0,1138 | 0,6815 |
| Efeito indireto via Ácido Aconítico | 0,0541 | 0,6200 |
| Total | -0,8433 | |
| | | |
| pH | | |
| Efeito direto sobre Cor ICUMSA | -0,0521 | 7,8531 |
| Efeito indireto via Brix | -0,0657 | 11,2688 |
| Efeito indireto via Pol | -0,0849 | 11,0447 |
| Efeito indireto via AR | 0,0108 | 3,2283 |
| Efeito indireto via ATR | -0,1033 | 11,3764 |
| Efeito indireto via Cu | -0,0687 | 11,7080 |
| Efeito indireto via Al | -0,2040 | 3,1549 |
| Efeito indireto via Mg | 0,0736 | 3,9174 |
| Efeito indireto via Ca | -0,0736 | 1,8604 |
| Efeito indireto via K | -0,1975 | 8,4772 |
| Efeito indireto via Compostos Fenólicos | -0,1023 | 0,5501 |
| Efeito indireto via Ácido Aconítico | 0,0504 | 0,5382 |
| Total | -0,8202 | |
| | | |
| Cu | | |
| Efeito direto sobre Cor ICUMSA | 0,0734 | 14,8024 |
| Efeito indireto via Brix | 0,0697 | 12,6767 |
| Efeito indireto via Pol | 0,0890 | 12,1559 |
| Efeito indireto via AR | -0,0109 | 3,2962 |
| Efeito indireto via ATR | 0,1088 | 12,6156 |
| Efeito indireto via pH | 0,0488 | 6,2114 |
| Efeito indireto via Al | 0,2237 | 3,7925 |
| Efeito indireto via Mg | -0,0775 | 4,3495 |
| Efeito indireto via Ca | 0,0672 | 1,5493 |
| Efeito indireto via K | 0,2081 | 9,4080 |
| Efeito indireto via Compostos Fenólicos | 0,1025 | 0,5525 |
| Efeito indireto via Ácido Aconítico | -0,0509 | 0,5476 |
| Total | 0,8556 | |

Continuação,,,

Tabela, continuação.

| Variáveis | | Efeito | VIF |
|---|---|---|---|
| Efeito direto sobre Cor ICUMSA | | 0,2596 | 5,6558 |
| Efeito indireto via Brix | Método 1 | 0,0670 | 11,7072 |
| Efeito indireto via Pol | | 0,0855 | 11,1988 |
| Efeito indireto via AR | k = 0,05 | -0,0109 | 3,3248 |
| Efeito indireto via ATR | | 0,1039 | 11,5171 |
| Efeito indireto via pH | | 0,0410 | 4,3805 |
| Efeito indireto via Cu | | 0,0633 | 9,9257 |

| | | |
|---|---|---|
| Efeito indireto via  Mg | -0,0644 | 2,9982 |
| Efeito indireto via Ca | 0,0340 | 0,3970 |
| Efeito indireto via K | 0,1703 | 6,3038 |
| Efeito indireto via Compostos Fenólicos | 0,0951 | 0,4758 |
| Efeito indireto via Ácido Aconítico | -0,053 | 0,5941 |
| Total | 0,8049 | |
| | | |
| Mg | | |
| Efeito direto sobre Cor ICUMSA | -0,0839 | 5,6453 |
| Efeito indireto via Brix | 0,0630 | 10,3413 |
| Efeito indireto via Pol | 0,0793 | 9,6477 |
| Efeito indireto via AR | -0,0091 | 2,2957 |
| Efeito indireto via ATR | 0,0975 | 10,1387 |
| Efeito indireto via pH | 0,0457 | 5,4494 |
| Efeito indireto via Cu | 0,0679 | 11,4047 |
| Efeito indireto via Al | 0,1991 | 3,0038 |
| Efeito indireto via Ca | 0,0541 | 1,0030 |
| Efeito indireto via K | 0,1811 | 7,1288 |
| Efeito indireto via Compostos Fenólicos | 0,0853 | 0,3829 |
| Efeito indireto via Ácido Aconítico | -0,0436 | 0,4015 |
| Total | 0,7319 | |
| | | |
| Ca | | |
| Efeito direto sobre Cor ICUMSA | 0,1048 | 4,1774 |
| Efeito indireto via Brix | 0,0346 | 3,1167 |
| Efeito indireto via Pol | 0,0433 | 2,8682 |
| Efeito indireto via AR | -0,0039 | 0,4305 |
| Efeito indireto via ATR | 0,0542 | 3,1269 |
| Efeito indireto via pH | 0,0366 | 3,4973 |
| Efeito indireto via Cu | 0,0471 | 5,4899 |
| Efeito indireto via Al | 0,0842 | 0,5376 |
| Efeito indireto via Mg | -0,0433 | 1,3554 |
| Efeito indireto via K | 0,1672 | 6,0729 |
| Efeito indireto via Compostos Fenólicos | 0,0541 | 0,1538 |
| Efeito indireto via Ácido Aconítico | -0,0392 | 0,3253 |
| Total | 0,544 | |

Continuação,,,

Tabela, continuação.

| Método 1 | | |
|---|---|---|
| k = 0,05 | | |
| Variáveis | Efeito | VIF |
| K | | |
| Efeito direto sobre Cor ICUMSA | 0,2173 | 11,369 |
| Efeito indireto via Brix | 0,0633 | 10,4585 |
| Efeito indireto via Pol | 0,0790 | 9,5795 |
| Efeito indireto via AR | -0,0084 | 1,9696 |
| Efeito indireto via ATR | 0,0978 | 10,1950 |
| Efeito indireto via pH | 0,0474 | 5,8556 |
| Efeito indireto via Cu | 0,0703 | 12,2491 |
| Efeito indireto via Al | 0,2035 | 3,1360 |
| Efeito indireto via Mg | -0,0700 | 3,5398 |
| Efeito indireto via Ca | 0,0806 | 2,2314 |
| Efeito indireto via Compostos Fenólicos | 0,0874 | 0,4018 |
| Efeito indireto via Ácido Aconítico | -0,0481 | 0,4893 |
| Total | 0,8312 | |
| Compostos Fenólicos | | |
| Efeito direto sobre Cor ICUMSA | 0,2010 | 2,3542 |
| Efeito indireto via Brix | 0,0403 | 4,2439 |
| Efeito indireto via Pol | 0,0574 | 5,0550 |
| Efeito indireto via AR | -0,0105 | 3,0738 |
| Efeito indireto via ATR | 0,0667 | 4,7481 |
| Efeito indireto via pH | 0,0265 | 1,8351 |
| Efeito indireto via Cu | 0,0374 | 3,4738 |
| Efeito indireto via Al | 0,1228 | 1,1430 |
| Efeito indireto via Mg | -0,0356 | 0,9182 |
| Efeito indireto via Ca | 0,0282 | 0,2730 |
| Efeito indireto via K | 0,0945 | 1,9405 |
| Efeito indireto via Ácido Aconítico | -0,0626 | 0,8287 |
| Total | 0,578 | |
| Ácido Aconítico | | |
| Efeito direto sobre Cor ICUMSA | -0,0963 | 2,1752 |
| Efeito indireto via Brix | 0,0420 | 4,5984 |
| Efeito indireto via Pol | 0,0554 | 4,6994 |
| Efeito indireto via AR | -0,0082 | 1,8695 |
| Efeito indireto via ATR | 0,0662 | 4,6755 |
| Efeito indireto via pH | 0,0273 | 1,9429 |
| Efeito indireto via Cu | 0,0388 | 3,7263 |
| Efeito indireto via Al | 0,1428 | 1,5448 |
| Efeito indireto via Mg | -0,0380 | 1,0420 |
| Efeito indireto via Ca | 0,0427 | 0,6248 |
| Efeito indireto via K | 0,1085 | 2,5572 |
| Efeito indireto via Compostos Fenólicos | 0,1306 | 0,8969 |
| Total | 0,5074 | |
| Coeficiente de determinação ($R^2$) | 0,7915 | |
| Efeito da variável residual | 0,4566 | |
| Determinante da matriz $X'X$ | $5,80x10^{-8}$ | |

## CONSIDERAÇÕES FINAIS

Este trabalho apresentou no capítulo 1 a técnica de análise de trilha para quantificar os efeitos diretos e indiretos em dados de cana-de-açúcar, nos estágios de cana-planta e cana-soca. Dentre os caracteres avaliados, em dois experimentos, número de colmos (*NC*) foi a variável que melhor se correlacionou com tonelada de colmos por hectare (*TCH*), demonstrando a possibilidade de obtenção de ganhos significativos por meio da seleção indireta para *TCH* via *NC*.

A análise de trilha possibilitou verificar que houve variação entre os experimentos, o que provavelmente se deve à origem diferenciada das famílias avaliadas. Assim, os resultados obtidos são pertinentes apenas para este estudo, havendo necessidade de avaliação de um maior número de experimentos.

No capítulo 2, foram comparados dois métodos alternativos de estimação dos coeficientes de trilha em presença de multicolinearidade, fazendo uso de dados em cana-soca, obtidos do programa de melhoramento da cana-de-açúcar da Universidade Federal de Viçosa reais obtidos do programa de melhoramento da cana-de-açúcar. O método baseado na regressão em crista e a exclusão de variáveis por componentes principais apresentaram resultados semelhantes na estimativa dos coeficientes de trilha, portanto, a escolha de um desses procedimentos depende do conhecimento e interesse do pesquisador. Com os resultados encontrados neste estudo, foi possível identificar os caracteres Al, K e compostos fenólicos como aqueles que melhor explicam a Cor do caldo. Para fins de melhoramento, a seleção indireta para Cor do caldo, por meio de índice de seleção envolvendo as variáveis Brix, Pol, AR, ATR, pH, Cu, Al, Mg, Ca, K, Compostos fenólicos e Ácido aconítico é recomendável.