**APLICAÇÃO DE TÉCNICAS ESTATÍSTICAS MULTIVARIADAS EM DADOS DE CUSTOS DA FASE DE PÓS-COLHEITA DO CAFÉ**

R. V. M. Santos[[1]](#footnote-1), H. D. Vieira[[2]](#footnote-2), F. M. Borém[[3]](#footnote-3), M. V. B. Prado[[4]](#footnote-4)

**RESUMO:** A escolha do modo de processamento do café é decisiva na rentabilidade da atividade cafeeira, e dependerá de diversos fatores. Assim, em decorrência da existência de tantas variáveis é comum o produtor questionar a viabilidade de determinados tipos de processamentos. Desse modo, o objetivo deste trabalho foi realizar um estudo dos principais fatores influenciadores no custo da pós-colheita do café. Quarenta e seis fazendas das regiões do Cerrado, Matas de Minas e Sul de Minas Gerais responderam a um questionário elaborado no sentido de possibilitar esta análise. A aplicação das técnicas multivariadas de análise de agrupamento, análise de fatores e análise de componentes principais, possibilitou concluir que as fazendas com maiores custos simulados foram as que possuíam as maiores porcentagens de produção de café via úmida.

Palavras-chave:agrupamento, componentes principais, análise de fatores.

**APPLICATION OF MULTIVARIATE STATISTICAL TECHNIQUES**

**FOR DATA ON THE COSTS OF THE POST-HARVEST PROCESSING OF COFFEE**

**ABSTRACT:** The choice of method of processing coffee is decisive on the profitability of the coffee activity, and will depend on several factors. Thus, due to the existence of many variables is common producer questioned the viability of certain types of processing. Thus, the objective of this study was to conduct a study of the major influencing factors in the cost of post-harvest coffee. Forty-six farms in the regions of the Cerrado, Matas de Minas and southern Minas Gerais answered a questionnaire in order to enable this analysis. The application of multivariate techniques of cluster analysis, factor analysis and principal component analysis, allowed us to conclude that farms with higher costs simulated were those with the highest percentages of wet coffee production.

Keywords: clustering, principal components, factor analysis.

**1 INTRODUÇÃO**

Em uma época em que o mercado de cafés especiais está se expandindo cada vez mais e o consumidor ficando mais exigente, elevar a qualidade da bebida é fundamental para atender o mercado e aumentar a renda do cafeicultor (Saath et al., 2010; Resende et al., 2011; Mendonça et al., 2007). Um dos fatores determinantes para isto é a pós-colheita (Abrahão et al., 2010). Borém (2008) afirma que grande parte das questões agronômicas da cafeicultura brasileira já estão muito bem evoluídas. Todavia, no momento em que o café é colhido e encaminhado para o processamento faltam informações de gestão de conhecimento tecnológico que garantam a manutenção da qualidade com baixos custos.

De acordo com Field et al. (2007) citado por Valente et al. (2011), a determinação dos custos é relevante para a este tipo de análise, sendo uma base para a tomada de decisão.

O Pós-Café consiste em um sistema de apoio à decisão voltado a análise do custo da pós-colheita do café, ou seja, emprega o conhecimento humano para resolver problemas que requerem a presença de um especialista nesta fase de processamento do café (Borém, 2008).

As técnicas multivariadas possibilitam avaliar um conjunto de características, considerando as correlações existentes, permitindo que inferências sobre o conjunto de variáveis sejam feitas em um nível de significância conhecido (Mardia et al., 1997).

Para a utilização da análise de variância multivariada, trabalhasse uma ampla gama de metodologias multivariadas possibilitando complementar os resultados obtidos e propiciar subsídios para utilização prática das informações resultantes (Benin et al., 2009).

Devido ao grande número de variáveis que influenciam os custos de processamentos do café em sua fase de pós-colheita, esse trabalho tem por objetivo estudar as principais variáveis influentes neste custo, por meio da utilização de um sistema de apoio à decisão (Pós-Café) e técnicas estatísticas multivariadas.

**2 MATERIAL E MÉTODOS**

**Aplicação de questionários aos produtores de café**

Foram aplicados questionários em 46 fazendas localizadas nas regiões: Cerrado, Matas de Minas e Sul de Minas Gerais. Escolheram-se estas regiões devido aos diversos tipos de processamentos na fase de pós-colheita do café utilizados nas fazendas. Dessa forma, o estudo das variáveis relacionadas aos custos da pós-colheita do café tornou-se mais completo.

**Simulações de programa computacional**

Os dados de cada questionário respondido nas 46 fazendas foram utilizados como entrada em simulações processadas no sistema de apoio à decisão para análise do custo da pós-colheita do café, denominado Pós-Café (Borém, 2008).

**Análise estatística**

Os softwares utilizados para realização das análises foram: Statistica e o R. Nesta etapa do trabalho abordaram-se técnicas estatísticas multivariadas para o estudo aprofundado dos dados dos questionários aplicados às 46 fazendas.

Análise de agrupamento das fazendas: foram trabalhadas variáveis quantitativas, por análise de agrupamento, na finalidade de classificá-las em grupos relativamente homogêneos e verificar as similaridades entre as mesmas (Mardia et al., 1997). Os dados foram padronizados e a distância escolhida foi a euclidiana média, devido à ausência de repetições nos tratamentos. O método de agrupamento escolhido foi o método da ligação completa, o qual tem tendência a formar grupos compactos.

Análise de fatores: utilizada para explicar o comportamento de um número relativamente grande de variáveis respostas, em termos de um pequeno número de fatores. Além disso, também foram estudadas as correlações entre as variáveis, principalmente em relação ao custo simulado da pós-colheita. Para a seleção do número de componentes optou-se pelo critério de Kaiser, citado por Mardia et al. (1997), ou seja, autovalores maiores que 1.

Análise de componentes principais: teve caráter confirmatório, tanto com objetivo de confirmação dos grupos de variáveis formados pela análise de agrupamento, quanto em relação às correlações identificadas na análise de fatores (Jackson, 1980).

**3 RESULTADOS E DISCUSSÃO**

**Análise estatística multivariada para o custo da fase de pós-colheita do café**

A Tabela 1 apresenta as respostas das 46 fazendas aos questionários aplicados para algumas das variáveis de entrada. Apesar de outras variáveis também terem sido utilizadas na simulação de custos do Pós-Café, estas foram selecionadas para a análise multivariada por influenciarem diretamente o custo de pós-colheita do café.

Tabela 1 - Dados de 46 fazendas entrevistadas para as variáveis de entrada selecionadas.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Fazenda | QS  (sacas) | QDC  (dias) | PCVU  (%) | CMO  (R$) | TER  (m2) | SEC  (litros) | CSIM  (R$) |
| 1 | 19000 | 80 | 30 | 30 | 18000 | 120000 | 6,98 |
| 2 | 10000 | 85 | 0 | 50 | 21000 | 0 | 4,08 |
| 3 | 6000 | 120 | 28,33 | 42 | 10000 | 0 | 8,66 |
| 4 | 4000 | 60 | 0 | 50 | 7000 | 39000 | 11,82 |
| 5 | 1700 | 100 | 0 | 50 | 3800 | 0 | 9,36 |
| 6 | 2100 | 60 | 0 | 30 | 5500 | 0 | 7,76 |
| 7 | 6000 | 90 | 35,83 | 50 | 6000 | 64000 | 17,72 |
| 8 | 9000 | 90 | 11 | 25 | 17900 | 51000 | 8,04 |
| 9 | 4000 | 120 | 50 | 45 | 7300 | 30000 | 11,17 |
| 10 | 900 | 45 | 0 | 50 | 3000 | 0 | 8,59 |
| 11 | 1500 | 60 | 0 | 40 | 4000 | 0 | 5,24 |
| 12 | 4500 | 70 | 0 | 45 | 9000 | 0 | 8,35 |
| 13 | 5000 | 80 | 0 | 40 | 8800 | 39000 | 9,61 |
| 14 | 6000 | 75 | 0 | 60 | 10300 | 0 | 10 |
| 15 | 13500 | 60 | 17 | 40 | 21000 | 60000 | 8,98 |
| 16 | 2300 | 70 | 0 | 50 | 9500 | 0 | 8,13 |
| 17 | 800 | 26 | 0 | 50 | 5000 | 0 | 9,32 |
| 18 | 8000 | 50 | 0 | 45 | 6000 | 0 | 10,2 |
| 19 | 1300 | 80 | 0 | 35 | 3000 | 0 | 7,01 |
| 20 | 8000 | 105 | 0 | 45 | 10600 | 45000 | 6,95 |
| 21 | 6500 | 45 | 0 | 50 | 10000 | 88000 | 8,55 |
| 22 | 8500 | 90 | 0 | 46 | 11000 | 30000 | 9,27 |
| 23 | 3500 | 95 | 0 | 30 | 12000 | 0 | 6,05 |
| 24 | 4700 | 80 | 0 | 30 | 10000 | 0 | 4,54 |
| 25 | 4000 | 90 | 0 | 40 | 6000 | 22000 | 5,25 |
| 26 | 18000 | 120 | 31 | 40 | 46600 | 123000 | 14,16 |
| 27 | 16000 | 100 | 35 | 40 | 20000 | 171000 | 9,41 |
| 28 | 6000 | 90 | 40 | 55 | 12000 | 60000 | 9,39 |
| 29 | 6000 | 80 | 0 | 40 | 10000 | 30000 | 10,85 |
| 30 | 38000 | 110 | 48 | 30 | 66000 | 150000 | 6,92 |
| 31 | 6000 | 72 | 56 | 45 | 27000 | 72000 | 11,69 |
| 32 | 35000 | 90 | 20 | 45 | 85000 | 126000 | 9,25 |
| 33 | 11000 | 60 | 40 | 45 | 20000 | 90000 | 9,36 |
| 34 | 1800 | 100 | 0 | 50 | 4500 | 0 | 15,24 |
| 35 | 3000 | 120 | 78 | 30 | 6600 | 22500 | 24,23 |
| 36 | 12200 | 45 | 65 | 40 | 47000 | 141000 | 14,77 |
| 37 | 4000 | 50 | 57,5 | 49 | 10500 | 65500 | 25,77 |
| 38 | 16500 | 140 | 30 | 51,5 | 29500 | 84000 | 9,83 |
| 39 | 27500 | 110 | 0 | 26 | 37300 | 210000 | 10,18 |
| 40 | 1804 | 75 | 56 | 40 | 6500 | 35000 | 15,19 |
| 41 | 12000 | 100 | 40 | 30 | 20000 | 90000 | 17,18 |
| 42 | 20000 | 100 | 47 | 42 | 27700 | 105000 | 11,29 |
| 43 | 10000 | 100 | 0 | 42 | 21400 | 36000 | 10,78 |
| 44 | 5312 | 100 | 55 | 75 | 10500 | 67500 | 12,57 |
| 45 | 8000 | 120 | 55 | 30 | 8000 | 69000 | 12,99 |
| 46 | 5000 | 90 | 60 | 30 | 3000 | 57500 | 18,07 |

As variáveis apresentadas na Tabela 1 possuem os seguintes significados:

* QS - Quantidade de sacas de café colhidas (sacas);
* QDC - Quantidade de dias necessários para a colheita do café (dias);
* PCVU - Porcentagem de café via úmida produzidos (%);
* CMO - Custo de mão de obra por dia de trabalho (R$);
* TER - Área de terreiro declarada (metros quadrados);
* SEC - Capacidade de secadores declarada (litros);
* CSIM - Custo simulado de pós-colheita por saca de café (R$).

As seis primeiras variáveis tem como respostas os dados declarados pelos responsáveis das fazendas entrevistadas. Já a variável número 7, a saber, CSIM, possui dados originados por intermédio de simulações de software realizadas na ferramenta Pós-Café.

Para os dados referentes aos custos da fase de pós-colheita do café contidos na Tabela 1 foram aplicadas as técnicas de análise multivariada: análise de agrupamento, análise de fatores e análise de componentes principais.

**Análise de agrupamento para os custos da fase de pós-colheita do café**

Na realização da análise de agrupamento, para dados padronizados, pelo método de ligação completa e distância euclidiana, o resultado foi o dendograma da Figura 1.



Figura 1 - Dendograma vertical dos dados de custos da fase de pós-colheita do café.

No dendograma elaborado verifica-se, tomando como base a distância 5,60, a formação de 4 grupos distintos: um grupo composto pela variável CMO; um grupo composto pelas variáveis CSIM e PCVU; um grupo composto pela variável QDC; e um grupo composto pelas variáveis SEC, TER e QS. Uma vez que estes grupos são formados pelas similaridades existentes entre as variáveis, cabe ressaltar que o CSIM e PCVU possuem elevada homogeneidade, ou seja, são variáveis correlacionadas.

**Análise de fatores para os custos da fase de pós-colheita do café**

Além de se realizar um estudo das correlações entre as variáveis, mediu-se a adequação dos dados através do teste KMO, o qual forneceu um valor de 0,61 e o *Bartlett Test* com valor de 2669,629 e nível de significância de p = 0,000. Através desses valores demonstra-se que a análise de fatores pôde ser conduzida, obtendo um grau de adequação próximo de razoável, conforme a classificação do KMO (Bezerra e Corrar, 2006).

Esta análise explicou o comportamento das sete variáveis respostas, em termos de dois fatores. As variáveis respostas foram agrupadas por meio de suas correlações.

Considerando as sete variáveis e analisando-se os resultados obtidos, pode-se verificar que o 1° e 2º autovalores são maiores que 1 e explicam 65,83% da variância, conforme apresentado pela Tabela 2. Os demais autovalores são menores que 1. Portanto, os dados serão resumidos pelo 1º e 2º componentes principais.

Tabela 2 - Extração de autovalores: utilizando o método das componentes principais.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Valor | Autovalor | % Variância Total | Autovalor Cumulativo | Cumulativo % |
| 1 | 2,984130 | 42,63043 | 2,984130 | 42,6304 |
| 2 | 1,624432 | 23,20617 | 4,608562 | 65,8366 |
| 3 | 0,983149 | 14,04498 | 5,591711 | 79,8816 |
| 4 | 0,796313 | 11,37590 | 6,388024 | 91,2575 |
| 5 | 0,286391 | 4,09130 | 6,674415 | 95,3488 |
| 6 | 0,261929 | 3,74184 | 6,936344 | 99,0906 |
| 7 | 0,063656 | 0,90937 | 7,000000 | 100,0000 |

A Figura 2 mostra os planos fatoriais para custos da fase de pós-colheita do café. Analisando-a, pode-se observar que as variáveis se agrupam por similaridades de explicação, ou seja, estão agrupadas por fatores. As variáveis que explicam ou representam melhor o fator 1 são aquelas que estiverem localizadas mais distantes da origem, em relação ao eixo x. As variáveis que estiverem localizadas próximas à origem do plano fatorial possuem baixa representatividade. Neste caso, as variáveis que melhor representam o fator 1, num prolongamento do eixo x até a origem dos eixos, são: QS, SEC e TER. Do mesmo modo, as variáveis que melhor representam o fator 2 são: CSIM e PCVU.

As distâncias de cada uma das variáveis analisadas são encontradas traçando-se segmentos de reta perpendiculares ao eixo considerado. Assim, as variáveis que estiverem localizadas mais distantes da origem, a partir da perpendicular traçada, terão maior influência sobre o fator analisado.



Figura 2 - Planos fatoriais para custos da fase de pós-colheita do café.

Também foram determinadas às correlações entre as variáveis, conforme a Tabela 3.

Tabela 3 - Correlações para custos da fase de pós-colheita do café.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | QS | QDC | PCVU | CMO | TER | SEC | CSIM |
| QS | 1,00 | 0,33 | 0,19 | -0,25 | 0,90 | 0,79 | -0,11 |
| QDC | 0,33 | 1,00 | 0,28 | -0,18 | 0,22 | 0,24 | 0,09 |
| PCVU | 0,19 | 0,28 | 1,00 | -0,07 | 0,23 | 0,44 | 0,65 |
| CMO | -0,25 | -0,18 | -0,07 | 1,00 | -0,15 | -0,22 | 0,03 |
| TER | 0,90 | 0,22 | 0,23 | -0,15 | 1,00 | 0,69 | -0,06 |
| SEC | 0,79 | 0,24 | 0,44 | -0,22 | 0,69 | 1,00 | 0,17 |
| CSIM | -0,11 | 0,09 | 0,65 | 0,03 | -0,06 | 0,17 | 1,00 |

Analisando-se os resultados da matriz de correlação (Tabela 3) entre as variáveis, observou-se que a maioria dos valores eram inferiores a 0,7 indicando uma correlação fraca entre as variáveis. Apenas entre as variáveis QS e TER; QS e SEC existiam valores de correlações fortes, concluindo-se uma interligação entre as variáveis citadas e que a análise das mesmas deveria ser realizada conjuntamente e não de forma isolada.

Todavia é importante notar que a variável mais correlacionada (0,65) ao custo simulado (CSIM) foi a porcentagem de café via úmida (PCVU).

**Análise de componentes principais para os custos da fase de pós-colheita do café**

Para esta análise também foram consideradas apenas as duas primeiras componentes principais, conforme Tabela 2.

O círculo de correlação unitário (Figura 3) foi utilizado como complemento para um maior esclarecimento das análises realizadas.

Analisando-se a Figura 3 observa-se que as variáveis que estiverem mais próximas ao círculo unitário, possuem uma maior contribuição em relação àquelas que estiverem mais afastadas. Também deve ser observado o ângulo formado entre duas variáveis, demonstrando maior ou menor correlação entre as mesmas. Assim, TER e SEC; e CSIM e PCVU contribuem fortemente na análise e possuem correlação significativa, devido ao pequeno ângulo formado entre estas variáveis.

Fazendo-se uma sobreposição da Figura 3 sobre o 1° plano fatorial (Figura 4), identificam-se visualmente quais as variáveis que estão relacionadas com os casos em estudo, ou seja, as fazendas. Como exemplo, podemos citar que para a variável CSIM os casos mais significativos são: 2, 24 (menores valores de custos); 35, 37 (maiores valores de custos).



Figura 3 - Círculo de correlação unitário para custos da fase de pós-colheita do café.



Figura 4 - Distribuição da nuvem de pontos por fazendas para custos da fase de pós-colheita.

Analisando-se o gráfico anterior, conclui-se que diversos casos contribuem na formação da combinação linear dos fatores 1 e 2, tornando-se muito difícil a identificação dos casos mais representativos, podendo-se citar: 30 e 32 para o fator 1; 35 e 37 para o fator 2.

Observa-se uma distribuição bastante uniforme das fazendas, ocorrendo uma maior concentração em torno da origem. Esses casos não apresentam representatividade significativa, em relação aos fatores 1 e 2. Também ocorrem algumas sobreposições, ou seja, algumas fazendas foram influenciadas pelos fatores 1 e 2 de forma muito parecida.

**4 CONCLUSÃO**

As fazendas podem ser agrupadas em relação às produções em sacas de café (QS), áreas de terreiro declaradas (TER) e capacidades de secadores declaradas (SEC). Além disso, as fazendas com maiores custos simulados (CSIM) são aquelas que possuem as maiores porcentagens de produção de café via úmida (PCVU).

O custo simulado por saca de café na fase de pós-colheita foi caracterizado pela independência quanto à influência de diversos fatores estudados. Como apenas a variável de porcentagem de produção de café via úmida apresentou correlação relativamente significativa em relação ao custo, conclui-se que a influência dos demais fatores ocorre de forma independente para as fazendas estudadas.

É importante ressaltar que o fato da porcentagem de produção de café via úmida estar correlacionada positivamente com o custo, não permite concluir que é sempre viável utilizar a metodologia via seca na fase de pós-colheita do café, uma vez que a mesma apresenta custos menores. Isso porque, esta decisão depende não somente do custo, mas também da receita.

A aplicação das técnicas multivariadas de análise de agrupamento, análise de fatores e análise de componentes principais, possibilitou complementaridade dos resultados obtidos.

**AGRADECIMENTOS**

Os autores agradecem à Faperj, CNPq, Capes e Nucoffee por financiarem este trabalho.

**5 REFERÊNCIAS**

ABRAHÃO, S. A.; PEREIRA, R. G. F. A.; DUARTE, S. M. S.; LIMA, A. R.; ALVARENGA, D. J.; FERREIRA, E. B. Compostos bioativos e atividade antioxidante do café (*Coffea arabica* L.). **Ciência e Agrotecnologia**, Lavras, v. 34, n. 2, p. 414-420, mar./abr., 2010.

BENIN, G.; SILVA, G. O.; PAGLIOSA, E. S.; LEMES, C.; SIGNORINI, A.; BECHE, E.; CAPELIN, M. A. Capacidade de combinação em genótipos de trigo estimada por meio de análise multivariada. **Pesquisa Agropecuária Brasileira**, Brasília, v.44, n.9, p.1145-1151, set., 2009.

BEZERRA, F. A.; CORRAR, L. J. Utilização da análise fatorial na identificação dos principais indicadores para avaliação do desempenho financeiro: uma aplicação nas empresas de seguros. **Revista Contabilidade & Finanças**, São Paulo, v. 1, n. 42, p. 50-62, set./dez., 2006.

BORÉM, F. M. Processamento do Café. In: Borém, F. M. **Pós-Colheita do Café**. 1. ed. Lavras: UFLA, 2008. Cap. 5, p. 127-158.

FIELD, F.; KIRCHAIN, R.; ROTH, R. Process cost modeling: strategic engineering and economic evaluation of materials technologies. **JOM Journal of the Minerals, Metals and Materials Society**, Warrendale, v. 59, p. 21-32, 2007.

JACKSON, J. E. Principal components and factor analysis: Part I - principal components. **Journal of Quality Technology**, New York, v.12, n.4, p. 201-213, out., 1980.

MARDIA, A. K. V.; KENT. J. T.; BIBBY, J.M. **Multivariate analysis**. London: Academic Press, 1997, 518p.

MENDONÇA, L. M. V. L.; PEREIRA, R. G. F. A.; MENDES, A. N. G.; BORÉM, F. M.; MARQUES, E. R. Composição química de grãos crus de cultivares de *Coffea arábica* L. suscetíveis e resistentes à *Hemileia vastatrix* Berg et Br. **Ciência e Agrotecnologia**, Lavras, v. 31, n. 2, p. 413-419, mar./abr., 2007.

RESENDE, O.; AFONSO, P. C.; CORREA, P. C.; SIQUEIRA, V. C. Qualidade do café conilon submetido à secagem em terreiro híbrido e de concreto. **Ciência e Agrotecnologia**, Lavras, v.35, n.2, p.327-335, mar./abr., 2011.

RIBEIRO, J. I.; MELO, A. L. P. **Guia prático para utilização do SAEG**. Viçosa: Folha, 2009. 45 p.

SAATH, R.; BORÉM, F. M.; ALVES, E.; TAVEIRA, J. H. S.; MEDICE, R.; CORADI, P. C. Microscopia Eletrônica de Varredura do Endosperma de Café (*Coffea arábica* L.) durante o processo de secagem. **Ciência e Agrotecnologia**, Lavras, v.34, n.1, p.196-203, jan./fev., 2010.

VALENTE D. S. M.; QUEIROZ D. M.; CORRÊA, P. C.; SILVA, L. C.; VALE, S. M. L. R. Um sistema de apoio à decisão para determinação de tarifas em unidades armazenadoras de grãos. **Engenharia Agrícola**, Jaboticabal, v. 31, n. 4, p. 735-744, Agosto 2011.

1. Instituto Federal do Espírito Santo/IFES - Coordenadoria de Informática - Cx. P. 527 - 29300-970 - Cachoeiro de Itapemirim - ES - Brasil - ravarmes@gmail.com [↑](#footnote-ref-1)
2. Universidade Estadual Norte Fluminense/UENF - Centro de Ciência e Tecnologias Agropecuárias/CCTA - Campos dos Goytacazes - RJ - Brasil - henrique@uenf.br [↑](#footnote-ref-2)
3. Universidade Federal de Lavras/UFLA - Departamento de Engenharia/DEG - Cx. P. 3037 - 37.200-000 Lavras - MG - Brasil - flavioborem@deg.ufla.br [↑](#footnote-ref-3)
4. Universidade Federal de Lavras/UFLA - Departamento de Estatística/DEX - Cx. P. 3037 - 37.200-000 Lavras - MG – Brasil - mari\_bernardes@yahoo.com.br [↑](#footnote-ref-4)