Aumento amostral via arquétipos na avaliação do potencial hídrico de espécies de eucalipto

Pórtya Piscitelli Cavalcanti [1] [4]
Carlos Tadeu dos Santos Dias [2] [4]
Patrícia Andressa de Ávila [3] [4]
José Leonardo de Moraes Gonçalves [3] [4]

Resumo

O objetivo desde trabalho é aumentar o tamanho amostral dos dados de potencial hídrico de diferentes espécies de eucalipto por meio de seus arquétipos, visando aumentar a precisão no procedimento de inferência estatística. Com essa finalidade, foram avaliadas seis espécies de eucaliptos: *E. grandis*, *E. urophylla*, *E. cloeziana*, *C. citriodora*, *E. camaldulensis* e *E. brassiana*, com quatro repetições e foram medidas duas variáveis-resposta: potencial hídrico mínimo e máximo. Então, para cada espécie, foram realizados dois aumentos sucessivos de um dado, aumentando o número de repetições de quatro para seis e, consequentemente, o tamanho amostral de 24 para 36 (aumento de 50% de n), sem alterar a distribuição de probabilidade da qual provém a amostra inicial e nem seus parâmetros. O aumento amostral corroborou para avaliar o efeito das seis espécies de eucalipto no potencial hídrico mínimo e máximo com maior precisão.

1 Introdução

Nas diversas áreas do conhecimento podem ser encontrados dados faltantes, ausentes ou incompletos (*missing data*) ou conjuntos de dados com menos observações do que o necessário para realizar o procedimento de inferência estatística com a precisão desejada. Nas ciências agronômicas, por exemplo, as unidades experimentais frequentemente são seres vivos, que podem vir a óbito (perda de parcelas), gerando dados desbalanceados. Além disso, existem casos em que se deseja utilizar o menor número de parcelas possível, por questões éticas, como em pesquisas com animais ou quando se tratam de amostras destrutivas.

Desta forma, com o intuito de completar a amostra, a técnica de aumento de dados consiste em aumentar de modo estatisticamente adequado um conjunto de dados

[1]LCE - ESALQ/USP. e-mail: *portya@usp.br*
[2]LCE - ESALQ/USP.
[3]LCF - ESALQ/USP.
[4]Adradecimentos à CAPES e ao CNPq pelo apoio financeiro.

observados, a fim de torná-lo mais propício para analisar [Tanner and Wong 1987]. O aumento amostral por meio de arquétipos é uma alternativa recente que vem sendo proposta [Cavalcanti 2016].

Arquétipos, na estatística, são os elementos extremos mais representativos de uma amostra ou população, a partir dos quais todos os outros podem ser escritos. A Análise de Arquétipos (AA) é uma técnica multivariada que visa, a princípio, reduzir a dimensão das observações, por meio de combinações convexas destas, proporcionando encontrar e selecionar seus arquétipos [Cutler and Breiman 1994]. Estes são selecionados por meio da minimização da soma de quadrados dos erros cometidos na reconstrução de cada observação como combinação dos arquétipos.

Portando, como os arquétipos são capazes de reescrever os elementos amostrais com um erro mínimo, é possível utilizá-los na geração de elementos não observados na amostra original [Cavalcanti 2016].

Assim, o objetivo desde trabalho é realizar o aumento amostral via arquétipos dos dados de potencial hídrico de diferentes espécies de eucalipto visando aumentar a precisão no procedimento de inferência estatística.

2 Materiais e Métodos

Foram avaliadas seis espécies de eucaliptos com diferentes níveis de tolerância à deficiência hídrica: *E. grandis* e *E. urophylla* (baixa tolerância), *E. cloeziana* e *C. citriodora* (tolerância intermediária), *E. camaldulensis* e *E. brassiana* (alta tolerância).

O potencial hídrico foliar dos eucaliptos foi medido aos 19 meses pós plantio em dezembro de 2017. Foram selecionadas quatro plantas de cada espécie com CAP (circunferência a altura do peito) média na área útil da parcela não-destrutiva, totalizando 24 árvores no experimento. As avaliações foram conduzidas nos períodos de antemanhã-AM ($\approx$ 4 h) e ao meio dia-MD ($\approx$ 12 h).

Assim, foram avaliados seis tratamentos (espécies de eucalipto) com quatro repetições (plantas) e foram medidas duas variáveis-resposta (potencial hídrico mínimo - AM, e máximo - MD), caracterizando um conjunto de dados bivariado.

Com os dados coletados, primeiramente verificou-se que a amostra provém de uma normal bivariada, por meio do teste de Royston, e, em seguida, foram estimados o seu vetor de médias e sua matriz de covariâncias amostral, que é não estruturada. Então, foi realizado o aumento amostral via arquétipos proposto por [Cavalcanti 2016].

O aumento de dados via arquétipos consiste, primeiramente, em aplicar a AA para encontrar um determinado número de arquétipos $\boldsymbol{z}_k \in {}_K\boldsymbol{Z}_p$, em que K é o número de arquétipos, p é o número de variáveis, n é o número de observações, $k = 1, \ldots, K$ e $K < n$, que sejam combinações lineares das observações

$$\boldsymbol{z}_k = \sum_{i=1}^{n} \boldsymbol{x}_i \beta_{ik} \tag{1}$$

em que $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ e β_{ik} são coeficientes que caracterizam a combinação convexa, ou seja, $\beta_{ik} \geq 0$ e $\sum_{i=1}^{n} \beta_{ik} = 1$. Desta forma, a AA minimiza a soma de quadrados dos resíduos (SQR) na reconstrução de cada observação como combinação dos arquétipos

$$SQR = \sum_{i=1}^{n} ||\boldsymbol{x}_i - \sum_{k=1}^{K} \alpha_{ik} \boldsymbol{z}_k||^2 \tag{2}$$

que também ocorre por combinação convexa, ou seja $\alpha_{ik} \geq 0$ e $\sum_{k=1}^{K} \alpha_{ik} = 1$.

Selecionados os arquétipos $\boldsymbol{z}_k$ e seus coeficientes α_{ik}, são sorteados coeficientes referentes a cada arquétipo, denominados α_{*k} em que $* \in \{1, 2, \ldots, n\}$ e, em seguida, os coeficientes sorteados são multiplicados pelos respectivos arquétipos, resultando nos dados não observados. O algoritmo utilizado nesse trabalho, conforme explicado em [Cavalcanti 2016], realiza esse aumento considerando uma restrição de combinação convexa dos arquétipos e consiste nos seguintes passos:

(i) Sortear um dos arquétipos $\boldsymbol{z}_k$, em que $k = 1, \ldots, K$;

(ii) Sortear um coeficiente referente ao primeiro arquétipo (α_{*k});

(iii) Calcular a diferença $d = 1 - \sum_{k=1}^{K} \alpha_{*k}$;

(iv) Se $d = 0$ e, consequentemente, $\sum_{k=1}^{K} \alpha_{*k} = 1$, encerrar o sorteio e zerar os coeficientes seguintes; se $d > 0$, sortear outro arquétipo dentre os restantes;

(v) Sortear um coeficiente referente ao arquétipo sorteado, desde que $\alpha_{*k} \leq d$;

(vi) Repetir os itens (iii) a (v) até chegar no último arquétipo.

(vii) Multiplicar os coeficientes sorteados pelos respectivos arquétipos.

Desta forma, foram realizados dois aumentos sucessivos de um dado para cada espécie, ou seja, aumentou-se um dado na amostra inicial e, na sequência, aumentou-se mais um dado na amostra já aumentada; alterando o número de repetições de quatro para seis e, consequentemente, o tamanho amostral de 24 para 36 (aumento de 50% de n).

É importante ressaltar que esses aumentos foram programados de modo a manter: a distribuição de probabilidade da variável aleatória, o vetor de médias e a matriz de covariâncias das observações de cada espécie; bem como a distribuição de probabilidade dos resíduos e a matriz de covariâncias residual, de acordo com os testes multivariados

de Royston, T^2 de Hotelling e M de Box que verificam, respectivamente, a normalidade multivariada, a igualdade entre vetores de médias e a igualdade entre matrizes de covariâncias.

Por fim, como os resíduos foram não correlacionados, cada variável resposta dos dados aumentados foi analisada univariadamente pela análise de variância e para comparação múltipla das médias das espécies foi utilizado o teste de Tukey. Em todas análises e testes realizados foi considerando um nível de 5% de significância.

3 Resultados e Discussões

O aumento de dados via arquétipos permitiu aumentar a amostra em 50% do seu tamanho inicial mantendo as características das observações. Na Figura1 pode-se notar que os dados aumentados apresentarem valores bem semelhantes aos dados observados e, em alguns casos, quase idênticos.

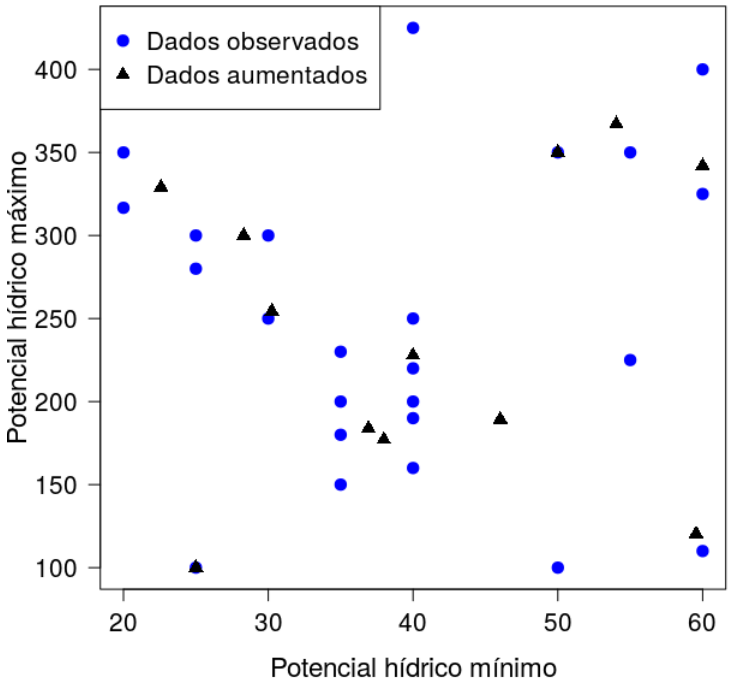

Figura 1: Dados observados e aumentados.

Além da análise visual, os resultados dos testes de Royston (valor-p = 0,8179) e M de Box (valor-p = 0,1014) confirmam que os dados aumentados possuem distribuição de probabilidade normal para os resíduos e matrizes de covariâncias residuais homogêneas, respectivamente.

Avaliando a correlação residual, verificou-se que esta foi muito próxima de zero ($r = 0,02$) e, portanto, não foi significativa (valor-p = 0,8911). Desta forma, prosseguiu-se com a análise univariada dos dados e verificou-se que ambas as variáveis-resposta foram

significativas, pois nos dois casos o valor-p foi menor que 0,0001. Sendo assim, as espécies de eucalipto apresentaram efeito significativo sobre o potencial hídrico mínimo e máximo.

A fim de comparar os potenciais hídricos mínimo e máximo entre as espécies de eucalipto, em seguida, as Figuras 2 e 3 apresentam os resultados do teste de Tukey para ambas variáveis-resposta.

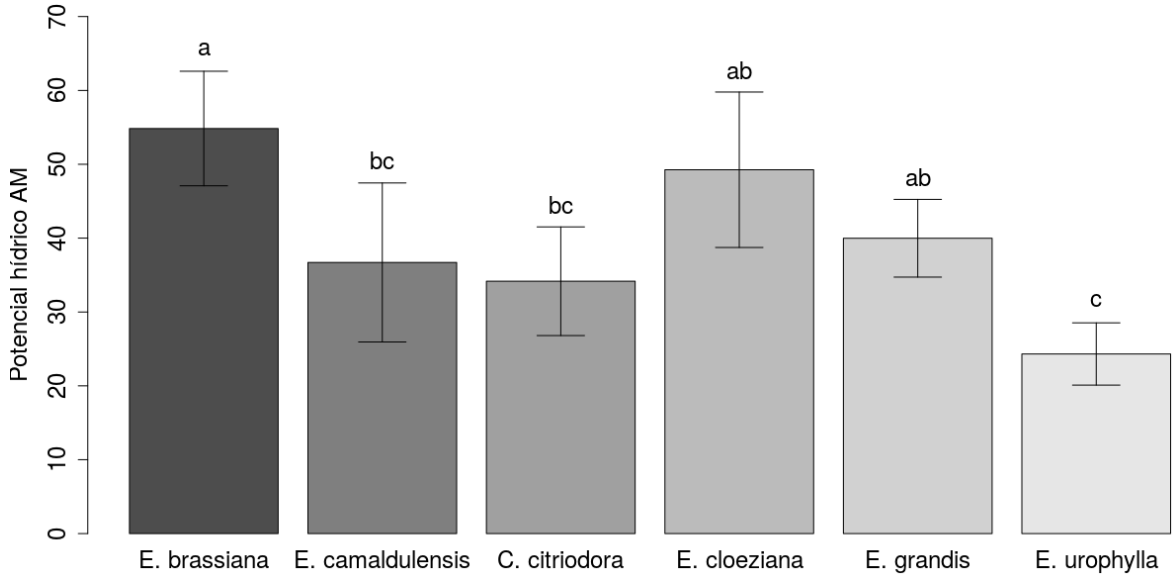

Figura 2: Médias de potencial hídrico mínimo das seis espécies de eucalipto.

Na Figura 2 verifica-se que o a espécie *E. brassiana* apresentou em média o maior potencial hídrico mínimo, sendo estatisticamente igual às espécies *E. cloeziana* e *E. grandis*, e a espécie *E. urophylla* apresentou o menor potencial hídrico mínimo, não diferindo estatisticamente das espécies *E. camaldulensis* e *C. citriodora*.

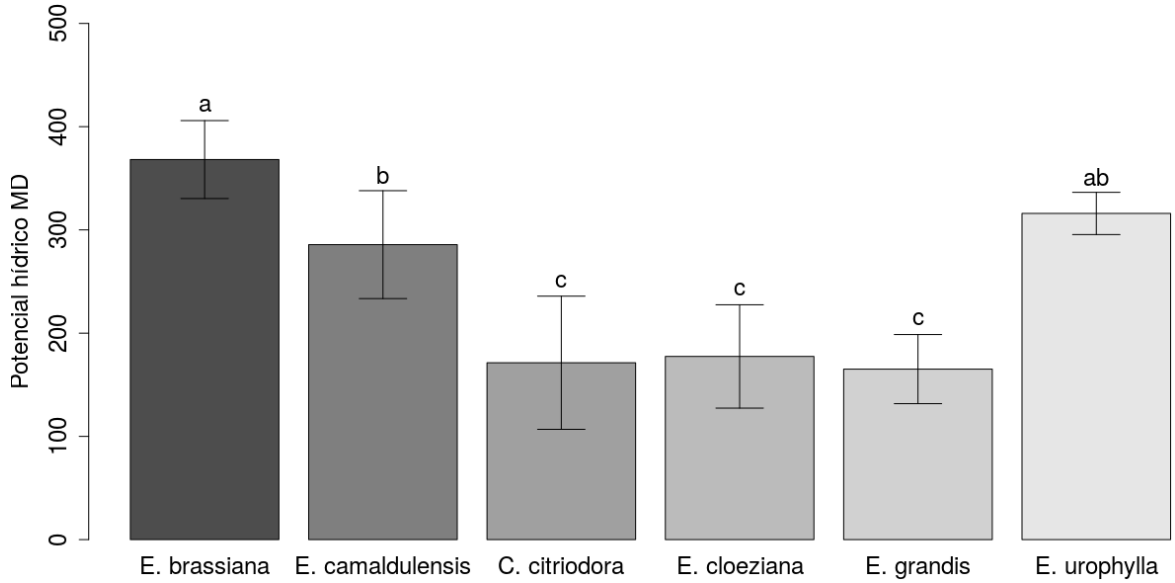

Figura 3: Médias de potencial hídrico máximo das seis espécies de eucalipto.

Avaliando o potencial hídrico máximo, em média a espécie *E. brassiana* apresentou maior valor de potencial hídrico e foi estatisticamente igual à espécie *E. urophylla*, já as espécies *C. citriodora*, *E. cloeziana* e *E. grandis* não diferiram estatisticamente e apresentaram as menores médias de potencial hídrico máximo.

4 Conclusão

O aumento amostral via arquétipos dos dados observados de potencial hídrico de espécies de eucalipto foi adequado, pois permitiu aumentar o tamanho amostral em 50% sem alterar a distribuição de probabilidade nem seus parâmetros. Assim, foi possível avaliar o efeito das seis espécies de eucalipto no potencial hídrico mínimo e máximo com maior precisão.

Referências

[Cavalcanti 2016]CAVALCANTI, P. P. *Proposta de algoritmos para aumento de dados via arquétipos*. 2016. Thesis (Master in Estatística Aplicada e Biometria), Universidade Federal de Alfenas - Unifal-MG, Alfenas, MG.

[Cutler and Breiman 1994]CUTLER, A.; BREIMAN, L. Archetypal analysis. *Technometrics*. v. 36, p. 338–347, 1994.

[Martins Júnior et al. 2015]MARTINS JÚNIOR, J. M. et al. A análise de arquétipos: uma revisão bibliográfica. *Revista Brasileira de Biometria*. 2015; 33:156–169.

[R Core Team 2018]R CORE TEAM. *R*: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2018. ISBN 3-900051-07-0.

[Tanner and Wong 1987]TANNER, M. A.; WONG, W. H. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* v. 82, p. 528–550, 1987.