Análise de dados em ciência do solo - como estão as pressuposições?

Cristiano Nunes Nesi[1], **Maurício Vicente Alves**[2], **Emanueli Maurilia Rebelatto**[3], **Susiane Chiamulera Migliavacca**[3]

Introdução

Apesar da importância dada à escolha do delineamento experimental e das análises estatísticas, frequentemente, os pesquisadores utilizam testes estatísticos inadequados devido à não verificação das pressuposições necessárias. Em consequência, surgem dificuldades tanto na análise dos dados, quanto na interpretação dos resultados obtidos, podendo assim, originar conclusões errôneas. A escolha do delineamento experimental adequado, bem como o emprego correto do modelo estatístico são de fundamental importância para a pesquisa. Entretanto, observa-se que muitos resultados de experimentos são publicados sem que as pressuposições das análises sejam mencionadas e, provavelmente, não sejam adequadamente verificadas. Quando as pressuposições são violadas a confiabilidade de todos os testes paramétricos ficam comprometidos, o que pode levar a falsas conclusões a respeito dos efeitos de tratamentos, conforme abordado há muito pela literatura especializada (Kempthorne, 1952; Cochran e Cox, 1957). Muitas vezes, apenas pela natureza dos dados, transformações são escolhidas por indicações genéricas propostas nos livros textos. A compreensão do não atendimento das pressuposições é também dificilmente investigada.

As pressuposições da análise de variância são de que os tratamentos, assim como os ambientes, são aditivos e que os erros experimentais se distribuem normal e independentemente com uma variância comum (Kempthorne, 1952; Cochran e Cox, 1957). Em análise de regressão, a escolha do modelo adequado deve considerar, além da lógica biológica, as estatísticas de ajuste. Esta escolha está diretamente relacionada à precisão das inferências e à conclusão sobre o fenômeno estudado. Há disponível na literatura especializada um conjunto de estatísticas que podem ser utilizadas como critério para avaliar a qualidade de ajuste em modelos de regressão (Ratkowsky, 1993; Azzalini, 1996; Zeviani, 2013).

De forma geral, as pressuposições das análises podem ser verificadas com as funções disponibilizadas no ambiente `R`, como evidenciaram Werner e Nesi (2017). Aliado a isto, hoje não se pode alegar limitação de software para uma adequada análise de dados.

O objetivo deste estudo foi apresentar um diagnóstico sobre a qualidade das análises dos dados apresentadas em artigos na área de ciência do solo.

Material e Métodos

Para realização deste estudo, foram avaliados 110 artigos publicados na Revista Brasileira de Ciência do Solo (Revista..., 2019) nos volumes 41 e 42, correspondentes aos anos de 2017 e 2018. Os artigos foram avaliados e classificados de acordo com a divisão

[1]UNOESC. e-mail: *cristiano.nesi@unoesc.edu.br*
[2]UNOESC. e-mail: *mauricio.alves@unoesc.edu.br*
[3]UNOESC. e-mail: *emanuelirebelatto17@gmail.com*
[3]UNOESC. e-mail: *susianemig@gmail.com*

da revista, o delineamento experimental, a técnica de análise dos dados apresentada na metodologia e nos resultados, software utilizado na análise, verificação das pressuposições e qualidade do ajuste. As observações foram tabuladas e são apresentadas estatíticas descritivas dos dados coletados. Além disso, foram anotados es descritos os principais problemas observados.

Resultados e Discussão

Na Tabela 1 são apresentadas as estatísticas gerais dos artigos avaliados. Nos casos em que empregou-se a pesquisa experimental, há predominância dos delineamentos inteiramente casualizado e em blocos casualizados em quase metado dos artigos avaliados. Quanto a técnica de análise, destacam-se também a análise de variância e de regressão como técnicas empregadas na análise dos dados, com relevante utilização de técnicas multivariadas. Com relação aos softwares, há um grande número de softwares relatados, com destaques para o R, SAS e SISVAR, embora mais de 20% dos artigos não informaram o software utilizado.

Table 1: Número de artigos e respectivos percentuais em função da divisão em que foi publicado, delineamento experimental, técnica de análise de dados e software utilizado em pesquisas publicadas na Revista Brasileira de Ciência do Solo, volumes 41 (2017) e 42 (2018).

	Divisão			Total
	Solo no espaço e no tempo	Processos e propriedades do solo	Uso e manejo do solo	
n°	17 (15,46%)	41 (37,27%)	52 (47,27%)	110
		Delineamento		
DIC	-	12	13	25 (22,73%)
DBC	-	9	19	28 (25,45%)
		Abordagem estatística		
ANAVA	1	22	30	53 (48,15%)
Regressão	2	6	14	22 (20,00%)
Geoestatística	3	2	2	7 (6,36%)
Multivariada	1	11	7	19 (17,27%)
		Programa		
R	5	11	4	20
SAS	1	2	15	18
SISVAR	-	5	3	8
Outros	1	20	15	36
Não Indicou	1	8	15	24

Os dados referentes à qualidade das análises realizadas são apresentados na Tabela 2. A verificação de pressuposições é relatada em aproximadamente 35% dos artigos, com destaque para a verificação apenas de normalidade dos resíduos e homocedasticidade na análise de variância. A verificação de aditividade do modelo não foi mencionada. Na análise de regressão, em torno de 23% informaram ter verificado qualidade do ajuste da regressão, embora não apresentem estatísticas de comparação para escolha do modelo utilizado. A escolha inadequada de um modelo de regressão pode modificar completamente as conclsões de um trabalho, assim como abordado por NESI (2018).

Table 2: Divisão em que foi publicado, pressuposições de análises e testes utilizados em pesquisas publicadas na Revista Brasileira de Ciência do Solo, volumes 41 (2017) e 42 (2018).

Aspecto	Divisão			Total
	Solo no espaço e no tempo	Processos e propriedades do solo	Uso e manejo do solo	
Homocedasticidade	-	6	7	13 (24,53%)
Normalidade	-	9	9	18 (33,96%)
Aditividade	-	-	-	0
Não verificou	1	14	19	34 (64,15%)
Qualidade do ajuste	0	1	4	5 (22,73%)
Tukey	1	9	15	25
Scott-Knott	-	7	5	12
Outros	-	4	9	13

Além das informações contempladas nas Tabelas 1 e 2, foram elencados problemas observados nas análises de variância e de regressão e no artigo como um todo, que serão descritos a seguir.

Principais problemas observados na análise de variância:

✓Não verificar as pressuposições;
✓Os autores relatam terem feito a verificação de normalidade dos resíduos, aplicam uma transformação mas não informam se verificaram sua efetividade;
✓Utilização de uma transformação genérica de acordo com a natureza da variável, principalmente a transformação raiz quadrada: $y* = \sqrt{y + c}$ em que $y*$ é a variável transformada, y a variável original e c uma constante.
✓Relato de verificação da normalidade dos dados, e não dos resíduos na análise de variância, o que aparenta desconhecimento dos autores em relação a diferença;
✓Teste de comparação de médias, sem qualquer informação sobre a análise de variância;
✓Mudança do nível de de significância de 5% para 25% sem correta justificativa, buscando evidenciar o efeito de tratamento;
✓Experimento e respectiva análise de variância realizada com apenas duas repetições;
✓Intervalos de confiança para comparar médias em mais que dois tratamentos;

✓Abordagem sobre a realização de análise de variância sem informar o delineamento utilizado no experimento;
✓Coeficiente de variação experimental acima de 70%, sem qualquer informação sobre a adequação da análise ou pressuposições;
✓Experimentos fatoriais em que não se desdobra a interação, apesar de informarem ser significativa;
✓Relatos de verificação da homocedasticidade com teste de Shapiro-Wilk;
✓Há utilização de softwares indiscriminadamente mesmo em caso de parcelas perdidas.

As alterações mais importantes originam-se quando a variância do erro experimental não é constante em todas as observações (heterocedasticidade). Algumas vezes isto ocorre porque certos tratamentos são mais erráticos (instáveis) em seus efeitos, enquanto outros, num mesmo experimento, apresentam resultados mais estáveis. Quando isto ocorre, as comparações que contêm tratamentos erráticos têm uma variância do erro maior que naquelas em que os tratamentos são mais estáveis. O mesmo problema pode aparecer quando os erros experimentais seguem uma distribuição assimétrica. Em tais distribuições, a variância do erro para um tratamento tende a ser uma função da média produzida pelo mesmo tratamento. Se a natureza da relação funcional é conhecida, uma alternativa é encontrar uma transformação que coloque os dados em uma escala tal que a variância do erro seja quase constante. Finalmente, há a suposição de que os erros entre as observações são independentes. Entretanto, sabe-se que os rendimentos de um cultivo em parcelas vizinhas tendem a estar positivamente correlacionadas, e em experimentos de laboratório as observações feitas por uma mesma pessoa tendem a apresentar idêntico tipo de correlação. Estas correlações podem alterar os testes de significância. Neste caso, o uso adequado da casualização introduz independência na atribuição dos tratamentos às unidades experimentais ou na atribuição da ordem em que se fazem as observações, de tal forma que os erros possam assim ser considerados efetivamente como independentes.

A aditividade ou o efeito aditivo do modelo permite a separação do efeito proveniente do tratamento daquele originado da variação do acaso, ou seja, permite distinguir os efeitos que influenciam no valor do resultado para uma determinada variável. Com a separação dos efeitos do tratamento e do acaso, é possível verificar se o resultado apresentado de uma observação é decorrente dos tratamentos aplicados, o que ocorre quando o efeito dos tratamentos apresenta-se muito superior ao efeito da variação do acaso. Se ao invés do efeito aditivo, por exemplo, houvesse um efeito multiplicativo, não haveria possibilidade de distinguir o efeito dos tratamentos e a variação do acaso.

Principais problemas na Análise de Regressão

✓Não abordam sobre critérios para a seleção de modelos;
✓Uso apenas do coeficiente de determinação (R^2) como medida de qualidade dos modelos;
✓Não apresentam a interpretação dos parâmetros nos modelos ajustados;
✓Não apresentam intervalos de confiança para as estimativas dos parâmetros;
✓Análises não descritas na metodologia que simplesmente 'surgem' nos resultados;

Em geral, a técnica usada para estimar os parâmetros em equações lineares ou não lineares é o método de mínimos quadrados ordinários que coincide com o método de máxima verossimilhança nos casos em que o modelo tem resposta normal independente e variância constante (Seber e Wild, 2003; Cordeiro et al., 2009). O estimador de mínimos quadrados

para parâmetros em modelos lineares são não viesados, normalmente distribuídos e de mínima variância. Quando estas pressuposições são satisfeitas, o critério de mínimos quadrados proporciona estimativas ótimas (Seber e Wild, 2003). Entretanto, para modelos não lineares as propriedades dos estimadores de mínimos quadrados são válidas apenas assintóticamente, ou seja, quando o tamanho da amostra aumenta para o infinito as propriedades do estimador aproximam-se das propriedades de mínimos quadrados para modelos lineares. Assim, percebe-se que todos os procedimentos inferenciais para modelos não lineares admitem suposição de adequada aproximação linear para fazerem uso de propriedades assintóticas. Desta forma, em modelos não lineares, quando se avalia se o modelo proposto proporciona uma boa descrição dos dados, a comparação do coeficiente de determinação e a análise de resíduos (Draper e Smith, 1998) podem ser insuficientes na seleção de modelos. A utilização dos critérios baseados nas estatísticas AIC e BIC e análise de resíduos é facilmente realizada, especialmente em ambiente `R` (Faraway, 2006).

Principais problemas gerais

✓Metodologia estatística pobremente descrita;
✓Não informar adequadamente os esquemas experimentais;
✓Não informar sobre o software utilizado, nem algoritmo de estimação;
✓Utilizar versões antigas de softwares, comprometendo a qualidade dos gráficos apresentados ou pode indicar a utilização de softwares sem as autorizações necessárias;

Considerações Finais

No geral há um aparente desleixo dos autores com realção à técnica de análise utilizada em suas pesquisas. Provavelmente há inferências realizadas sem que se tenha certeza da adequação da análise.

Sugere-se que se dê ênfase em equipes interdisciplinares em projetos de pesquisa, dando espaço aos profissionais habilitados para a realização das análises.

Referências Bibliográficas

AZZALINI, A. Statistical Inference - Based on the likelihood . London: Chapman & Hall, 341p, 1996.

BANZATTO, D. A.; KRONKA, S. N. *Experimentação Agrícola.* Jaboticabal: FUNEP. 2006. 237p.

COCHRAN, W.G.; COX, G.M. Experimental designs. 2.ed. London, John Wiley, 1957. 611p.

CORDEIRO, G.M.; PRUDENTE, A.A.; DEMÉTRIO, C.G.B. Uma revisão dos modelos normais não-lineares. **Revista Brasileira de Biometria**, 27(3): 360-393, 2009.

FARAWAY, J.J. Extending the Linear Model with R Generalized Linear, Mixed Effects and Nonparametric Regression Models. Chapman & Hall/CRC. 2006.

FERREIRA, E. B.; CAVALCANTI, P. P. Função em código R para analisar experimentos em DIC simples, em uma só rodada. In: 54ª Reunião da Região Brasileira da Sociedade Internatcional de Biometria, 13º Simpósio de Estatística Aplicada à Experimentação Agronômica, 2009, São Carlos. *Programas e resumos...* São Carlos, SP: UFSCar, 2009. p. 1-5.

KEMPTHORNE, O. (1952). The Design and Analysis of Experiments. Wiley ed.631.

NESI, C.N.. Modelo quadrático ou de Mitscherlich?. In: XII Reunião Sul Brasileira de Ciência do Solo, 2018, Xanxerê. Anais da XII Reunião Sul Brasileira de Ciência do Solo. Xanxerê: UNOESC, 2018.

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

RATKOWSKY, D.A. **Nonlinear regression modeling**, New York: M. Dekker, 1993.

REVISTA BRASILEIRA DE CIÊNCIA DO SOLO. versão On-line ISSN 1806-9657. http://www.scielo.br.

SEBER, G.A.F.; Wild, C.J. **Nonlinear regression**, New York: J. Wiley, 2003.

WERNER, S.S.; NESI, C.N.. Análise de variância com verificação das pressuposições - uma abordagem com o ambiente R. 2017. In: 62ª Reunião da Região Brasileira da Sociedade Internatcional de Biometria, 17º Simpósio de Estatística Aplicada à Experimentação Agronômica, 2017, Lavras. *Minicurso...* Lavras, MG: UFLA, 2017.

ZEVIANI, W. Parametrizações interpretáveis em modelos não lineares. Tese (Doutorado em Estatística e Experimentação Agropecuária). 146 p. Lavras: UFLA, 2013.