Ciência Rural, Santa Maria, v.53:5, e20220180, 2023

http://doi.org/10.1590/0103-8478cr20220180

ISSNe 1678-4596
CROP PRODUCTION

Sample size affects the precision of the analysis of variance in experiments with cauliflower seedlings

Karina Chertok Bittencourt[1] Marcos Toebe[1*] Rafael Rodrigues de Souza[2] Stella Bonorino Pazetto[3] Iris Cristina Datsch Toebe[4]

[1]Departamento de Ciências Agronômicas e Ambientais, Universidade Federal de Santa Maria (UFSM), 98400-000, Frederico Westphalen, RS, Brasil. E-mail: m.toebe@gmail.com. *Corresponding author.
[2]Departamento de Fitotecnia, Universidade Federal de Santa Maria (UFSM), Santa Maria, RS, Brasil.
[3]Universidade Federal do Pampa (UNIPAMPA), Itaqui, RS, Brasil.
[4]Programa de Pós-graduação em Informática na Educação, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brasil.

ABSTRACT: *This study verified whether sample size would affect the precision of the analysis of variance in experiments with cauliflower seedlings. An experiment was carried out where the number of leaves and shoot, root and total length were measured. For each variable, resamplings with repositions were performed in sample scenarios of 1, 2, ..., 100 seedlings per experimental unit, and the sample size was defined for the variance components through Schumacher models and maximum curvature points. The mean squares of the analysis of variance suffer direct interference from the number of sampled seedlings. The sampling of 16 seedlings per experimental unit is enough to estimate the analysis of variance reliably, promoting satisfactory precision gains compared to the sampling of only one seedling per experimental unit.*

Key words: *Brassica oleracea, horticulture, experimental planning, precision gain.*

O tamanho de amostra afeta a precisão da análise de variância em experimentos com mudas de couve-flor

RESUMO: *Este estudo verificou se o tamanho de amostra afetaria a precisão da análise de variância em experimentos com mudas de couve-flor. Um experimento foi conduzido onde o número de folhas, comprimento de parte aérea, raiz e total foram mensurados. Para cada variável, reamostragens com reposição foram realizadas em cenários amostrais de 1, 2, ..., 100 mudas por unidade experimental e o tamanho de amostra foi definido para os componentes de variância por meio de modelos de Schumacher e pontos de máxima curvatura. Os quadrados médios da análise de variância sofrem interferência direta do número de mudas amostradas. A amostragem de 16 mudas por unidade experimental é suficiente para estimar a análise de variância de forma confiável, promovendo satisfatórios ganhos de precisão ao comparar-se com a amostragem de apenas uma muda por unidade experimental.*

Palavras-chave: *Brassica oleracea, horticultura, planejamento experimental, ganho de precisão.*

In a previous research, four methods based on the maximum curvature point were compared to determine the optimal sample size per experimental unit to estimate the overall experimental mean of cauliflower (*Brassica oleracea* L. var. botrytis) seedlings (BITTENCOURT et al., 2022), where a reduction in the 95% confidence interval width ($CI_{95\%}$) of the statistic was verified as sample size increased, up to a stabilization point. Thus, the methods that reported values closer to the stabilization point of the curve were chosen, once precision gain up from this point would no longer be enough to justify increasing the number of sampled plants (CARGNELUTTI FILHO et al., 2018; SOUZA et al., 2022). That example highlighted the importance of quantifying precision gain when defining sample size, which would not only facilitate the decision on the number of plants to be sampled per experimental unit but would also guarantee a minimum acceptable precision to the results. However, the previous approach focused only on the overall experimental mean without exploring other components of the analysis of variance.

The analysis of variance is widely performed to summarize data in experiments with experimental designs (WELHAM et al., 2015).

Received 03.29.22 Approved 05.27.22 Returned by the author 07.12.22
CR-2022-0180.R1
Editors: Leandro Souza da Silva Alessandro Dal'Col Lúcio

Nonetheless, in order to find actual significant differences through the F test that follows it, mean squares must be estimated reliably, reducing the probability of type I and II errors (ANDERSON et al., 2017). For this, sample size plays a crucial role, as verified by SOUZA et al. (2022) for soybean crop, based on its impact for estimating other statistics in experiments performed with crotalaria and maize (TOEBE et al., 2018; CARGNELUTTI FILHO & TOEBE, 2021). Therefore, considering that studies connecting sample size and the precision gain of the analysis of variance have not been reported in the literature for horticultural crops such as cauliflower, this study verified whether sample size would affect the precision of the analysis of variance in experiments with cauliflower seedlings.

The experiment was carried out at the Federal University of Pampa (UNIPAMPA), Itaqui, Rio Grande do Sul, Brazil. Cauliflower cultivar Teresopolis Gigante was sown using three substrate mixtures (50% Mecplant® + 50% Carolina Padrão®, 75% Mecplant® + 25% rice husk, and 75% Carolina Padrão® + 25% rice husk), and trays with 72 and 128 cells, forming a 3x2 two-factor scheme, in a completely randomized design with four repetitions. Seedlings were kept in a greenhouse for a period of thirty days. During the sampling, twenty seedlings were randomly collected from each experimental unit, considering the sample numbers used in cauliflower experiments (THOMSON et al., 2013; TEMPESTA et al., 2019; COSTA et al., 2020). Then, the following traits were measured: a) Number of Leaves (NL) in units; b) Shoot Length (SL), from neck to leaflet insertion, in cm; c) Root Length (RL), from neck to root apex, in cm; and d) Total Length (TL), as the sum of SL and RL, in cm. Other experiments with 1, 2, …, 100 seedlings per experimental unit were simulated using bootstrap resampling, with 10,000 resamples with reposition (EFRON, 1979).

The statistical analyses were performed using native functions and packages from R software (R DEVELOPMENT CORE TEAM, 2022). First, the database was stratified into experimental units, and in each sample size, an analysis of variance was performed through the following mathematical model: $Y_{ijk} = m + T_i + S_j + (TS)_{ij} + \varepsilon_{ijk}$, where Y_{ijk} is the value observed in the response variable in plot *ijk*, *m* is the overall mean, T_i is the fixed effect of level *i* (i = 1 and 2) of the tray-cell-size factor, S_j is the fixed effect of level *j* (j = 1, 2, 3) of the substrate factor, $(TS)_{ij}$ is the interaction fixed effect of level *i* of the tray-cell-size factor with level *j* of the substrate factor and ε_{ijk} is the experimental error effect. Thereafter, the mean squares of T_i, S_j, $(TS)_{ij}$, and ε_{ijk} were extracted in the sample scenarios per experimental unit. This process was carried out using *sample*() and *aov*() functions.

Resamplings for each planned sample scenario were subjected to descriptive analysis defining minimum values, percentiles of 2.5, mean, percentiles of 97.5, and maximum values. The 95% confidence interval width ($CI_{95\%}$) was estimated as the difference between percentiles of 97.5 and percentiles of 2.5. Posteriorly, the precision gain criterion was estimated in percentage, assuming that the greater the $CI_{95\%}$, the lower the precision of the analysis-of-variance mean squares' estimates (SOUZA et al., 2022). Thus, the sample size of one seedling per experimental unit (CI_1) was taken as a reference, where the $CI_{95\%}$ is maximum and the precision is minimum. The following formula was used to estimate precision gain:

$$PG = 100 - \left(\frac{CI_i}{CI_1}\right) * 100$$

where CI_i is the 95% confidence interval width, obtained from the sample sizes of 2, 3, ..., 100 seedlings per experimental unit [for further information *vide* CARGNELUTTI FILHO et al. (2018) and SOUZA et al. (2022)].

Finally, the precision gain was fitted using *nls*() function through Schumacher's model (SCHUMACHER, 1939): $PG_i = a \times \exp(\beta \times n^{-1}) + \varepsilon$, where PG_i is the i^{th} precision gain observation per statistic, in each *n* sample size, *α* and *β* are parameters of the model, *exp* is the exponential function and ε_i is the error of random effect. A maximum curvature point was defined over the fitted models through the perpendicular distances' method (SILVA & LIMA, 2017), as recommended by BITTENCOURT et al. (2022) for cauliflower, using the *maxcurv*() function from the soilphysics package (SILVA & LIMA, 2015).

The variance components fluctuated in response to the variation of the number of seedlings sampled per experimental unit, also varying for each specific trait (Figure 1). In all cases, $CI_{95\%}$ tends to reduce gradually as the number of sampled seedlings is increased, which means estimates become more accurate (TOEBE et al., 2018; BITTENCOURT et al.,

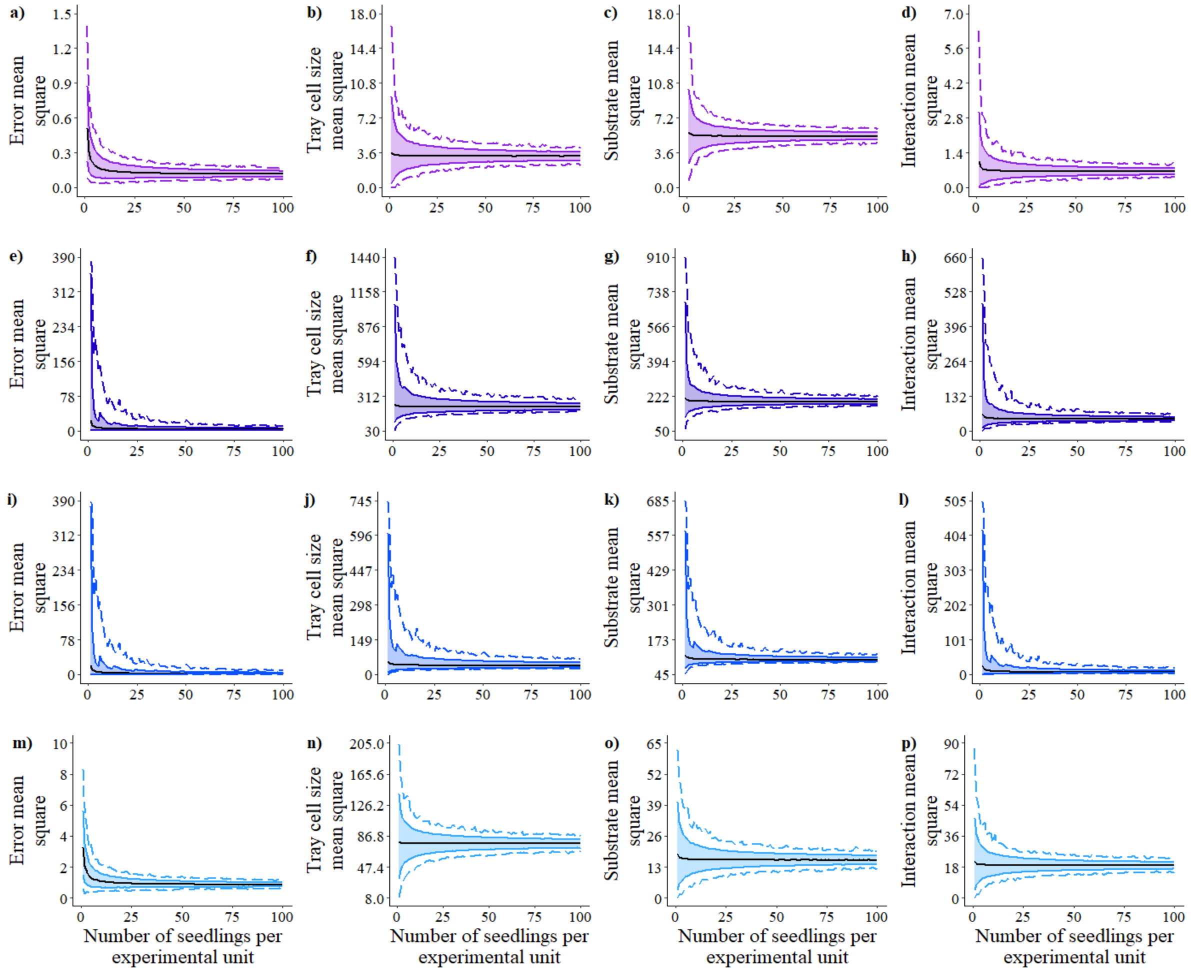

Figure 1 - Minimum, 2.5 percentile, mean, 97.5 percentile and maximum values of the mean squares of the error, tray cell size, substrate, and tray cell size × substrate interaction in the number of leaves (a, b, c, and d), total length (e, f, g, and h), shoot length (i, j, k, and l), and root length (m, n, o, and p) of cauliflower seedlings.

2022; SOUZA et al., 2022). Conversely, small sample sizes (≤ 5 seedlings per experimental unit) result in greater $CI_{95\%}$, making the mean squares estimates more biased. These results are similar to the ones observed by SOUZA et al. (2022) when analyzing the response of variance components in soybean.

From this response, it was observed that the precision of the analysis-of-variance mean squares was increased as sample size increased, establishing a direct relationship between result reliability and the number of seedlings used for data collection, especially considering the influence of the analysis of variance in the determination of significant differences between treatments. In general, the sufficient sample sizes for obtaining reliable estimates of the analysis of variance varied from 13 to 16 cauliflower seedlings per experimental unit, with precision gains oscillating from ≥ 76.52% to ≤ 93.42%, depending on the variance component and trait analyzed (Table 1 and figure 2). These values were obtained through the parametrization of precise Schumacher models (SCHUMACHER, 1939),

Table 1 - Coefficient of determination (R^2), root mean square error (RMSE), and d index of the Schumacher models, precision gains, and sample sizes for the analysis of variance of the number of leaves (NL), shoot length (SL), root length (RL), and total length (TL) of cauliflower seedlings.

Trait	Statistic*	Schumacher model	R^2	RMSE	d	Precision gain (%)	Sample size
NL	EMS	$PG_i = 92.7268 \times \mathit{exp}(-2.0653 \times n^{-1})$	0.98	1.70	0.99	80.34	15
NL	TMS	$PG_i = 90.8951 \times \mathit{exp}(-2.6275 \times n^{-1})$	0.99	1.59	0.99	76.82	16
NL	SMS	$PG_i = 90.9115 \times \mathit{exp}(-2.7168 \times n^{-1})$	0.99	1.43	0.99	76.55	16
NL	IMS	$PG_i = 92.1433 \times \mathit{exp}(-2.2303 \times n^{-1})$	0.98	1.65	0.99	79.22	15
TL	EMS	$PG_i = 101.0929 \times \mathit{exp}(-0.9950 \times n^{-1})$	0.79	4.69	0.93	93.11	13
TL	TMS	$PG_i = 95.2857 \times \mathit{exp}(-1.4891 \times n^{-1})$	0.94	2.61	0.98	85.09	14
TL	SMS	$PG_i = 97.0627 \times \mathit{exp}(-1.3535 \times n^{-1})$	0.93	2.96	0.98	87.37	13
TL	IMS	$PG_i = 98.4280 \times \mathit{exp}(-1.2088 \times n^{-1})$	0.89	3.52	0.97	89.40	13
SL	EMS	$PG_i = 101.2845 \times \mathit{exp}(-0.9737 \times n^{-1})$	0.78	4.83	0.93	93.42	13
SL	TMS	$PG_i = 97.7196 \times \mathit{exp}(-1.1904 \times n^{-1})$	0.88	3.64	0.96	88.85	13
SL	SMS	$PG_i = 98.8328 \times \mathit{exp}(-1.1430 \times n^{-1})$	0.87	3.86	0.96	90.14	13
SL	IMS	$PG_i = 100.1760 \times \mathit{exp}(-1.0428 \times n^{-1})$	0.82	4.39	0.94	91.96	13
RL	EMS	$PG_i = 94.5164 \times \mathit{exp}(-1.9269 \times n^{-1})$	0.98	1.66	0.99	82.45	15
RL	TMS	$PG_i = 90.8445 \times \mathit{exp}(-2.6214 \times n^{-1})$	0.99	1.54	0.99	76.79	16
RL	SMS	$PG_i = 90.6974 \times \mathit{exp}(-2.6674 \times n^{-1})$	0.99	1.55	0.99	76.52	16
RL	IMS	$PG_i = 91.0810 \times \mathit{exp}(-2.5692 \times n^{-1})$	0.99	1.52	0.99	77.17	16

* EMS: error mean square; TMS: tray cell size mean square; SMS: substrate mean square; IMS: interaction mean square.

with coefficients of determination (R^2) ≥ 0.78, root mean square error (RMSE) oscillating from 1.43 to 4.83, and d index ≥ 0.93. Furthermore, in sample sizes ≤ 3, a considerable precision gain is observed every time there is an increase in the number of sampled seedlings. This response remains until the sampling number reaches 10 seedlings per experimental unit, up from where precision gain starts becoming lower and lower, until finally reaching the maximum curvature point, that is, the ideal sample size for each trait and variance component.

In that perspective, considering all traits and variance components jointly, the minimum sampling number of 16 seedlings per experimental unit can be recommended as sufficient to make accurate mean square estimates for the analysis of variance of experiments with cauliflower seedlings, corroborating the results obtained by BITTENCOURT et al. (2022). They suggested the sampling of at least 15 cauliflower seedlings per experimental unit to estimate the overall experimental mean. The collection of greater samples normally demands more resources and manpower that are not justified by the little precision gain obtained (TOEBE et al., 2015), and in some cases, oversampling may even result in greater variations between experimental units that can inflate the error mean square (SOUZA et al., 2022). This harms the detection of significant differences between treatments due to the occurrence of type II error (ANDERSON et al., 2017). Importantly, the practical results here obtained should be applied cautiously in cauliflower seedlings' experiments with experimental designs, and should not be used for other horticultural crops without performing preliminary studies, serving only as a support to researchers that conduct experiments with other species from the Brassicaceae family.

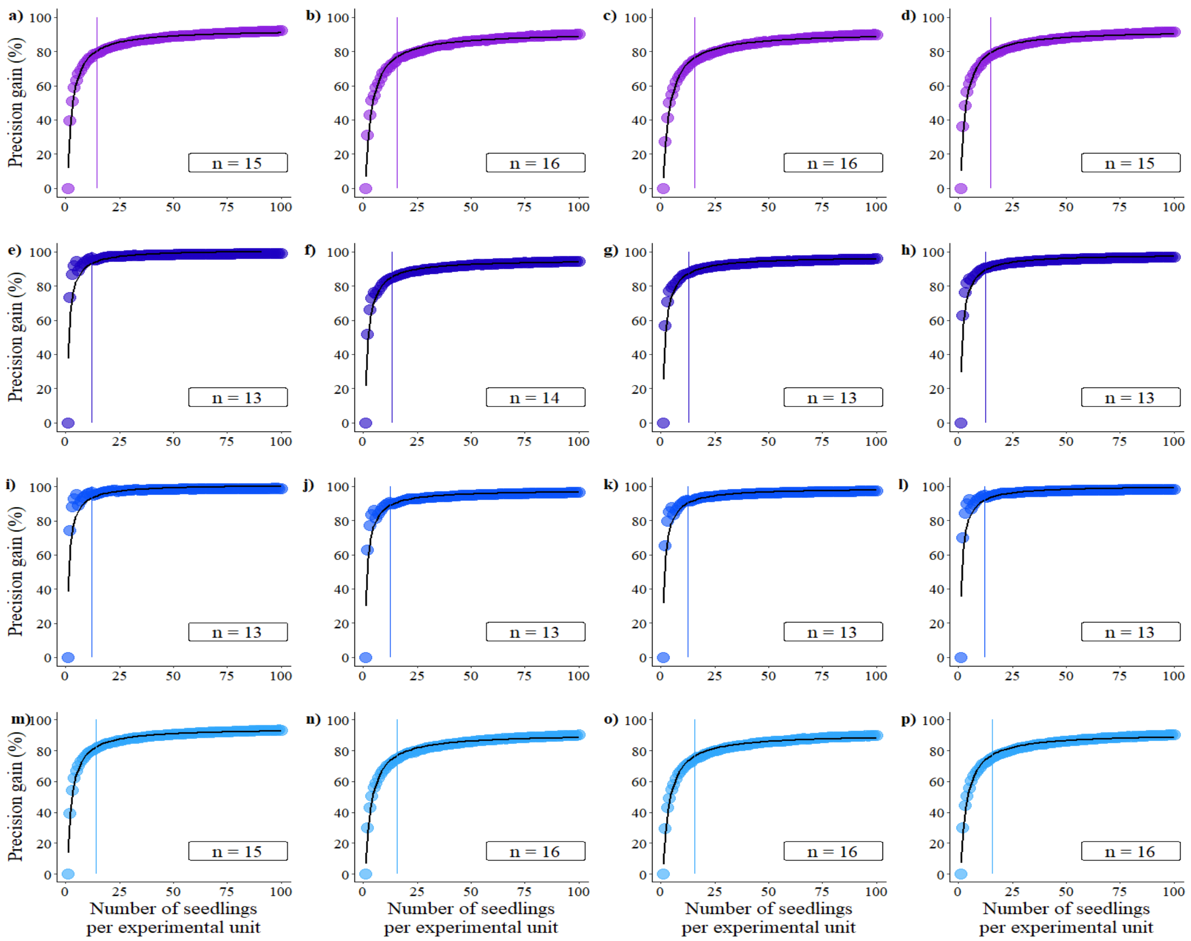

Figure 2 - Sample size determination via Schumacher model and maximum curvature points for estimating the mean squares of the error, tray cell size, substrate, and tray cell size × substrate interaction in the number of leaves (a, b, c, and d), total length (e, f, g, and h), shoot length (i, j, k, and l), and root length (m, n, o, and p) of cauliflower seedlings.

ACKNOWLEDGEMENTS

We thank the Universidade Federal do Pampa (UNIPAMPA) for infrastructure, the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Process 313827/2021-4 and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil - Finance code 001, for finantial support.

DECLARATION OF CONFLICT OF INTERESTS

We have no conflict of interest to declare.

AUTHORS' CONTRIBUTIONS

Conceptualization: KCB and RRS. Data acquisition: KCB and SBP. Design of methodology and data analysis: RRS and MT. Supervision and coordination: MT and ICDT. KCB and RRS prepared the draft of the manuscript. All authors critically revised the manuscript and approved of the final version.

REFERENCES

ANDERSON, S. F. et al. Sample-size planning for more accurate statistical power: a method adjusting sample effect sizes for publication bias and uncertainty. **Psychological Science**, v.28, p.1547–1562, 2017. Available from: <https://journals.sagepub.com/doi/10.1177/0956797617723724>. Accessed: Feb. 11, 2022. doi: 10.1177/0956797617723724.

BITTENCOURT, K. C. et al. What is the best way to define sample size for cauliflower seedlings? **Ciência Rural**, v.52, e20210747, 2022. Available from: <https://www.scielo.br/j/cr/a/w7pyKzKSLQ8zLbmG9qcF7mz/?lang=en>. Acessed: May 17, 2022. doi: 10.1590/0103-8478cr20210747.

CARGNELUTTI FILHO, A. et al. Number of leaves for modelling the leaf area of velvet bean according to leaf dimensions. **Revista de**

Ciências Agroveterinárias, v.17, p.571–578, 2018. Available from: <http://dx.doi.org/10.5965/223811711732018571>. Accessed: Jan. 22, 2022. doi: 10.5965/223811711732018571.

CARGNELUTTI FILHO, A.; TOEBE, M. Sample size for principal component analysis in corn. **Pesquisa Agropecuária Brasileira**, v.56, e02510, 2021. Available from: <http://dx.doi.org/10.5965/223811711732018571>. Accessed: Mar. 12, 2022. doi: 10.1590/S1678-3921.pab2021.v56.02510.

COSTA, L.F. et al. Cauliflower growth and yield in a hydroponic system with brackish water. **Revista Caatinga**, v.33, p.1060–1070, 2020. Available from: <http://dx.doi.org/10.1590/1983-21252020v33n421rc>. Accessed: Feb. 23, 2022. doi: 10.1590/1983-21252020v33n421rc.

EFRON, B. Bootstrap methods: another look at the jackknife. **Annals of Statistic**, v.7, p.1–26, 1979. Available from: <https://doi.org/10.1214/aos/1176344552>. Accessed: Feb. 18, 2022. doi: 10.1214/aos/1176344552.

R DEVELOPMENT CORE TEAM. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

SCHUMACHER, F. X. A new growth curve and its application to timber yield studies. **Journal of forestry**, v.37, p.819–820, 1939.

SILVA, A.R. da; LIMA, R.P. soilphysics: an R package to determine soil preconsolidation pressure. **Computers and Geosciences**, v.84, p.54–60, 2015. Available from: <https://doi.org/10.1016/j.cageo.2015.08.008>. Accessed: Feb. 20, 2022. doi: 10.1016/j.cageo.2015.08.008.

SILVA, A.R. da; LIMA, R.P. Determination of maximum curvature point with the R package soilphysics. **International Journal of Current Research**, v.9, p.45241–45245, 2017.

SOUZA, R.R. de. et al. Soybean yield variability per plant in subtropical climate: sample size definition and prediction models for precision statistics. **European Journal of Agronomy**, v.136, 126489, 2022. Available from: <https://doi.org/10.1016/j.eja.2022.126489>. Accessed: Mar. 15, 2022. doi: 10.1016/j.eja.2022.126489.

TEMPESTA, M. et al. Optimization of nitrogen nutrition of cauliflower intercropped with clover and in rotation with lettuce. **Scientia Horticulturae**, v.246, p.734–740, 2019. Available from: <https://doi.org/10.1016/j.scienta.2018.11.020>. Accessed: Jan. 25, 2022. doi: 10.1016/j.scienta.2018.11.020.

THOMSON, G. et al. Effects of elevated carbon dioxide and soil nitrogen on growth of two leafy Brassica vegetables. **New Zealand Journal of Crop and Horticultural Science**, v.41, p.69-77, 2013. Available from: <https://doi.org/10.1080/01140671.2013.772905>. Accessed: Jan. 23, 2022. doi:10.1080/01140671.2013.772905.

TOEBE, M. et al. Sample dimensioning for estimating coefficients of correlation in maize hybrids, harvests and precision levels. **Bragantia**, v.74, p.16–24, 2015. Available from: <https://doi.org/10.1590/1678-4499.0324>. Accessed: Jan. 23, 2022. doi:10.1590/1678-4499.0324.

TOEBE, M. et al. Sample size for estimating mean and coefficient of variation in species of crotalarias. **Anais da Academia Brasileira de Ciências**, v.90, p.1705–1715, 2018. Available from: <https://doi.org/10.1590/0001-3765201820170813>. Accessed: Jan. 23, 2022. doi: 10.1590/0001-3765201820170813.

WELHAM, S.J. et al. **Statistical methods in biology**: **Design and analysis of experiments and regression**. Boca Raton: CRC Press, 2015. 608p.