Universidade Federal Rural do Semi-Árido
Pró-Reitoria de Pesquisa e Pós-Graduação
https://periodicos.ufersa.edu.br/index.php/caatinga

ISSN 0100-316X (impresso)
ISSN 1983-2125 (online)
http://dx.doi.org/10.1590/1983-21252023v36n123rc

Sample size to evaluate the degree of multicollinearity in rye morphological traits

Tamanho de amostra para avaliação do grau de multicolinearidade em caracteres morfológicos de centeio

Ismael Mario Márcio Neu[1]*, Alberto Cargnelutti Filho[2], Marcos Toebe[3], Fernanda Carini[1], Rafael Vieira Pezzini[1], Daniela Lixinski Silveira[1]

[1]Postgraduate Program in Agronomy, Universidade Federal de Santa Maria, Santa Maria, RS, Brazil. [2]Department of Plant Sciences, Universidade Federal de Santa Maria, Santa Maria, RS, Brazil. [3]Department of Agronomic and Environmental Sciences, Universidade Federal de Santa Maria, Frederico Westphalen, RS, Brazil.

ABSTRACT - Investigation of multicollinearity allows parameters in multivariate analysis to be estimated with higher precision and with biological interpretation. In order to generate reliable estimates of the degree of multicollinearity, it is necessary to use appropriate sample size. Thus, the objectives of this study were to determine the sample size (number of plants) necessary to estimate the indicators of the degree of multicollinearity - condition number (CN), correlation matrix determinant (DET), and variance inflation factor (VIF) - in morphological traits of rye and to verify the variability of the sample size between the indicators. Five and three uniformity trials were conducted with the cultivars BRS Progresso and Temprano, respectively. Eight morphological traits were evaluated in 780 plants in eight trials. For each trial, 22 cases were selected among the 28 formed by the combination of eight traits, taken six by six, totaling 176 cases. In each case, 197 sample sizes were planned (20, 25, 30, ..., 1,000 plants) and in each size 2,000 resampling procedures with replacement were performed, CN, DET, and VIF were determined and the average among 2,000 estimates was calculated. For each case and indicator (CN, DET, and VIF), the sample size was determined through three models: modified maximum curvature method and linear and quadratic segmented models with plateau response. There is variability between sample sizes between indicators, with larger sample sizes required for DET, followed by CN and VIF, in that order, with at least 180, 116 and 85 plants, respectively.

Keywords: Correlation. Multivariate analysis. Sampling design. *Secale cereale* L.

RESUMO - A investigação da multicolinearidade permite que parâmetros em análises multivariadas sejam estimados com maior precisão e com interpretação biológica. Para ter confiabilidade nas estimativas do grau de multicolinearidade, é necessário utilizar adequado tamanho de amostra. Assim, os objetivos deste trabalho foram determinar o tamanho de amostra (número de plantas) necessário para a estimação dos indicadores do grau de multicolinearidade - número de condição (NC), determinante da matriz de correlação (DET) e fator de inflação da variância (FIV) - em caracteres morfológicos de centeio e verificar a variabilidade do tamanho de amostra entre os indicadores. Foram conduzidos cinco e três ensaios de uniformidade com as cultivares BRS Progresso e Temprano, respectivamente. Foram avaliados oito caracteres morfológicos em 780 plantas em oito ensaios. Para cada ensaio, foram selecionados 22 casos entre os 28 formados pela combinação de oito caracteres, tomados seis a seis, totalizando 176 casos. Para cada caso, foram planejados 197 tamanhos de amostra (20, 25, 30, ..., 1.000 plantas) e para cada tamanho foram realizadas 2.000 reamostragens, com reposição, determinados o NC, DET e FIV e calculada a média das 2.000 estimativas. Após, para cada caso e indicador, foi determinado o tamanho de amostra, por meio de três modelos: método da máxima curvatura modificado e modelos linear e quadrático segmentados com resposta em platô. Há variabilidade entre os tamanhos de amostra entre os indicadores, com necessidade de maiores tamanhos de amostra para DET, seguido de NC e FIV, nessa ordem, com no mínimo de 180, 116 e 85 plantas, respectivamente.

Palavras-chave: Correlação. Análise multivariada. Dimensionamento amostral. *Secale cereale* L.

Conflict of interest: The authors declare no conflict of interest related to the publication of this manuscript.

Received for publication in: May 19, 2021.
Accepted in: September 2, 2022.

***Corresponding author:**
<ismaelmmneu@hotmail.com>

INTRODUCTION

Rye (*Secale cereale* L.) belongs to the Poaceae family, with important use of its grains in human and animal diet, as a soil cover crop (SAPIRSTEIN; BUSHUK, 2016) and as forage crop (BAIER, 1994), with early supply of fodder at the end of autumn (PAULINO; CARVALHO, 2004), a time when other winter forage cereals are not yet at the ideal point for grazing. It is a crop with interesting characteristics to integrate crop rotation systems. It has high resistance to diseases and drought, tolerance to sandy and acidic soils (MORRISON, 2016), assists in the maintenance of soil water content (BASCHE et al., 2016) and exerts allelopathic or retarding effect on the germination of spontaneous plants (ABOU CHEHADE et al., 2021).

Breeding strategies can be obtained by knowing the correlation between crop characteristics (LAIDIG et al., 2017) and, in the selection process, univariate

and multivariate statistical techniques can be used as auxiliary tools. For the estimates of the parameters of the analysis to be reliable, it is necessary to assess the degree of multicollinearity between the predictor traits. Multicollinearity can be interpreted as the strong relationship between predictors and affects the precision with which coefficients are estimated (GUJARATI; PORTER, 2011; MONTGOMERY; PECK; VINNING, 2012). Inadequate interpretation of the parameters in canonical correlation analysis (ALVES; CARGNELUTTI FILHO; BURIN, 2017) as well as results with no biological meaning and estimates with no interpretation in path analysis (TOEBE; CARGNELUTTI FILHO, 2013) have been observed in studies conducted in the presence of multicollinearity.

Given the importance of the diagnosis of multicollinearity, it needs to be accurately estimated, which can be achieved using adequate sample size. The determination of sample size for agronomic characteristics has been carried out in studies with rye (BANDEIRA et al., 2018a; 2018b) and showy rattlepod (TOEBE et al., 2017a), as well as in the estimation of the correlation between traits of maize (OLIVOTO et al., 2017a) and parameters in path analysis in cherry tomato (SARI et al., 2018). In these studies, larger sample sizes promote greater precision, with reduced gain above the sample size determined. For rye, no studies determining the sample size necessary for the diagnosis of multicollinearity were found. In a study with rye crop, the diagnosis of multicollinearity was made with 128 observations (NOURAEIN, 2019), whereas in other crops, such as wheat (JANMOHAMMADI; SABAGHNIA; NOURAEIN, 2014), maize (OLIVOTO et al., 2017a; 2017b), showy rattlepod (TOEBE et al., 2017a), cherry tomato (SARI et al., 2018) and sunflower (FOLLMANN et al., 2019), the diagnosis was made with 45 to 1,180 observations. Therefore, the diagnosis of multicollinearity has been performed with different sample sizes, which generates estimates of lower or higher precision.

Some inferences have been made regarding sample size in the diagnosis of the degree of multicollinearity in maize traits (OLIVOTO et al., 2017a), as well as investigations regarding the interference of multicollinearity in path analysis in maize (TOEBE; CARGNELUTTI FILHO, 2013) and cherry tomato (SARI et al., 2018). Additionally, Olivoto et al. (2017a) and Sari et al. (2018) pointed out that insufficient sample sizes incorrectly estimate the degree of multicollinearity. However, these studies did not determine the appropriate sample size for estimating multicollinearity in rye traits.

Given the varied number of observations used in the diagnosis of multicollinearity and the existence of inferences made for the need to use larger sample sizes, this study was conducted. It is assumed that it is possible to determine the sufficient sample size (number of plants) for the diagnosis of the degree of multicollinearity and that this size differs between the indicators condition number, determinant and variance inflation factor. Thus, the objectives of this study were to determine the sample size (number of plants) necessary to determine the indicators of the degree of multicollinearity - condition number (CN), determinant (DET) and variance inflation factor (VIF) - in morphological traits of rye and to assess the variability of sample size between the indicators.

MATERIAL AND METHODS

Eight uniformity trials were conducted with rye crop (*Secale cereale* L.), consisting of five sowing times with the cultivar BRS Progresso (T1, T2, T3, T4 and T5) and three sowing times with the cultivar Temprano (T6, T7 and T8) in the winter crop season of 2016. These trials were conducted in an experimental area located in Santa Maria - RS (29º42' S, 53º49' W and 95 m altitude). According to Köppen's classification, the climate of the region is classified as Cfa - Humid subtropical climate, with hot summers and no defined dry season (ALVARES et al., 2013). The soil of the region is classified as *Argissolo Vermelho distrófico arênico* (Ultisol) (SANTOS et al., 2018).

The experimental area was homogeneously prepared and soil fertility was corrected with the application of 500 kg ha^{-1} of fertilizer (5-20-20 NPK formulation). Two rye cultivars were sown: BRS Progresso, intended for grain production; and Temprano, intended for soil cover and as forage plant. The seeds of each cultivar were sown broadcast in an area of 320 m² (20 m × 16 m) in the first sowing time, whereas in the other sowing times, each cultivar was sowned in an area of 375 m² (25 m × 15 m).

The sowing times were planned to meet the recommendation of planting from March to July (BAIER, 1994). For both cultivars and at all sowing times, a density of 455 seeds m^{-2} was used. Top-dressing fertilization was performed when the plants were between the stages of three and four developed leaves, using 25 kg ha^{-1} of nitrogen. The other cultural practices were carried out according to the need and to the management recommendations for rye crop (BAIER, 1994).

In each uniformity trial, 100 plants at physiological maturity were randomly collected, except for trials four and eight. In these trials, 90 plants were evaluated, corresponding to the cultivar BRS Progresso in the fourth sowing time and the cultivar Temprano in the third sowing time. In each plant, the following morphological traits were evaluated: number of stems $plant^{-1}$ (NSP = main stem + tillers); number of nodes $plant^{-1}$ (NNP = sum of the number of nodes of the stems); number of nodes $stem^{-1}$ (NNS = NNP/NSP); plant stem length, in cm (PSL = average length of stems); plant peduncle length, in cm (PPL = average length of stem peduncles); plant ear length, in cm (PEL = average length of ears); main stem height, in cm (MSH); and plant stem height, in cm (PSH = average height of the stems). PPL was defined as the stem portion between the last node and the ear insertion in the stem; PEL as the portion between the ear insertion in the stem and the last spikelet; and MSH and PSH as the portion between the base of the plant and the last spikelet. In this study, the data of plants of each trial were considered as the master sample.

For each trial, 28 cases were planned, obtained by combining eight traits taken six by six (Table 1). In each case, with the data from the master sample, the degree of multicollinearity was estimated by the indicators condition number (CN), correlation matrix determinant (DET), and variance inflation factor (VIF). CN was obtained by the relationship between the highest eigenvalue (λ_{max}) and the lowest eigenvalue (λ_{min}) of the correlation matrix ($CN = \lambda_{max}/\lambda_{min}$) (GUJARATI; PORTER, 2011) and classified as weak ($CN \leq 100$), moderate to strong ($100 < CN \leq 1{,}000$) and severe multicollinearity ($CN > 1{,}000$) (MONTGOMERY; PECK; VINNING, 2012). Problems due to multicollinearity may exist for DET lower than 0.00001 (FIELD, 2009) and VIF_j greater than or equal to ten, where $VIF_j = 1/(1 - R_j^2)$, where R_j^2 is the multiple coefficient of determination of a given variable with the other explanatory variables (GUJARATI; PORTER, 2011). CN and DET are indicators with interpretation for all variables, while VIF has the advantage of informing the variance inflation for each variable, and the highest VIF value was considered in this study.

Table 1. Traits combined in each case obtained by combining eight morphological traits of rye (*Secale cereale* L.) and the respective degree of multicollinearity (condition number - CN) of the master sample for each trial (two cultivars at different sowing times), evaluated in the 2016 season, Santa Maria, RS, Brazil.

C[1]	Traits[2]	Degree of multicollinearity - Condition number (CN)							
		T1[3]	T2	T3	T4	T5	T6	T7	T8
	---------- All eight traits ----------								
*	All traits	2.1×10^{16}	9.7×10^{16}	1.7×10^{17}	5.6×10^{16}	4.0×10^{16}	6.0×10^{16}	3.6×10^{16}	1.5×10^{16}
	---------- Eight traits combined seven by seven – C(8, 7) ----------								
*	x1;x2;x3;x4;x5;x6;x7	1.8×10^{17}	5.5×10^{16}	2.6×10^{17}	1.2×10^{17}	1.5×10^{17}	2.6×10^{16}	1.4×10^{16}	1.9×10^{16}
*	x1;x2;x3;x4;x5;x6;x8	1.8×10^{16}	2.5×10^{16}	1.0×10^{17}	2.8×10^{18}	2.1×10^{16}	6.9×10^{17}	3.6×10^{17}	1.1×10^{16}
*	x1;x2;x3;x4;x5;x7;x8	1.4×10^{17}	1.5×10^{16}	3.5×10^{16}	7.1×10^{16}	3.1×10^{16}	6.3×10^{16}	3.5×10^{17}	1.5×10^{16}
*	x1;x2;x3;x4;x6;x7;x8	6.4×10^{02}	1.5×10^{03}	9.2×10^{02}	1.2×10^{03}	7.6×10^{02}	5.9×10^{02}	7.2×10^{02}	1.0×10^{03}
*	x1;x2;x3;x5;x6;x7;x8	6.5×10^{02}	1.4×10^{03}	9.3×10^{02}	1.3×10^{03}	7.8×10^{02}	5.7×10^{02}	7.4×10^{02}	1.2×10^{03}
*	x1;x2;x4;x5;x6;x7;x8	5.4×10^{02}	1.3×10^{03}	8.1×10^{02}	1.2×10^{03}	6.9×10^{02}	5.3×10^{02}	6.4×10^{02}	1.1×10^{03}
*	x1;x3;x4;x5;x6;x7;x8	4.8×10^{02}	1.3×10^{03}	7.2×10^{02}	1.1×10^{03}	6.4×10^{02}	4.8×10^{02}	5.4×10^{02}	8.8×10^{02}
*	x2;x3;x4;x5;x6;x7;x8	3.6×10^{17}	6.0×10^{16}	2.1×10^{16}	1.5×10^{16}	1.1×10^{16}	1.4×10^{16}	1.4×10^{17}	1.5×10^{16}
	---------- Eight traits combined six by six – C(8, 6) ----------								
1*	x1;x2;x3;x4;x5;x6	2.3×10^{17}	2.3×10^{16}	9.2×10^{16}	9.0×10^{15}	1.7×10^{17}	1.4×10^{16}	4.8×10^{16}	2.2×10^{16}
2*	x1;x2;x3;x4;x5;x7	1.9×10^{17}	1.5×10^{16}	3.4×10^{17}	2.6×10^{16}	9.1×10^{16}	3.8×10^{16}	5.8×10^{16}	1.0×10^{16}
3*	x1;x2;x3;x4;x5;x8	2.2×10^{16}	1.5×10^{17}	1.9×10^{17}	1.7×10^{17}	1.3×10^{16}	4.0×10^{16}	1.7×10^{16}	4.6×10^{16}
4	x1;x2;x3;x4;x6;x7	72.9	35.9	58.7	142.1	71.2	87.7	44.9	27.2
5	x1;x2;x3;x4;x6;x8	117.5	283.1	235.3	207.2	205.7	117.5	239.8	220.8
6	x1;x2;x3;x4;x7;x8	72.4	36.2	59.3	146.5	73.2	85.2	44.9	27.7
7	x1;x2;x3;x5;x6;x7	269.5	233.4	464.1	373.3	627.8	385.2	285.5	504.3
8	x1;x2;x3;x5;x6;x8	262.9	322.5	444.6	367.3	608.0	355.5	272.4	495.4
9	x1;x2;x3;x5;x7;x8	268.6	237.4	468.3	374.8	635.5	378.4	285.2	502.2
10	x1;x2;x3;x6;x7;x8	603.5	1,402.0	885.1	1,151.2	675.5	538.2	674.7	967.6
11	x1;x2;x4;x5;x6;x7	65.5	26.8	26.8	135.1	59.6	58.8	19.5	24.4
12	x1;x2;x4;x5;x6;x8	110.9	240.6	219.3	207.6	189.6	100.7	221.1	241.8
13	x1;x2;x4;x5;x7;x8	65.5	27.5	27.1	138.2	61.7	58.2	19.8	24.6
14	x1;x2;x4;x6;x7;x8	475.5	1,261.9	704.1	984.3	620.5	465.8	538.9	816.6
15	x1;x2;x5;x6;x7;x8	470.2	1,132.4	746.5	1065.4	580.6	450.3	550.0	1,000.5
16*	x1;x3;x4;x5;x6;x7	1.1×10^{17}	3.1×10^{16}	9.2×10^{15}	5.3×10^{16}	2.9×10^{16}	3.6×10^{16}	6.6×10^{16}	1.3×10^{16}
17*	x1;x3;x4;x5;x6;x8	6.8×10^{16}	8.3×10^{15}	2.0×10^{16}	6.0×10^{16}	1.7×10^{16}	2.7×10^{16}	4.2×10^{16}	2.3×10^{16}
18*	x1;x3;x4;x5;x7;x8	1.4×10^{16}	1.3×10^{16}	9.0×10^{15}	8.0×10^{16}	9.8×10^{15}	2.7×10^{16}	7.1×10^{16}	2.1×10^{16}

[1]Cases. [2]Traits: main stem height (x1), plant stem height (x2), plant stem length (x3), plant ear length (x4), plant peduncle length (x5), number of stems plant^{-1} (x6), number of nodes stem^{-1} (x7) and number of nodes plant^{-1} (x8). [3]Trials conducted in five and three sowing times with the cultivars BRS Progresso (T1, T2, T3, T4 and T5) and Temprano (T6, T7 and T8), respectively. *Cases disregarded in the present study, due to the high degree of multicollinearity.

Table 1. Continuation.

C[1]	Traits[2]	Degree of multicollinearity - Condition number (CN)							
		T1[3]	T2	T3	T4	T5	T6	T7	T8
19	x1;x3;x4;x6;x7;x8	512.1	1,175.8	721.9	884.0	563.8	470.0	541.1	880.3
20	x1;x3;x5;x6;x7;x8	523.2	1,072.1	711.2	933.4	553.4	447.6	555.2	1,029.2
21	x1;x4;x5;x6;x7;x8	447.6	1,039.5	616.6	834.8	511.1	436.0	519.1	957.3
22	x2;x3;x4;x5;x6;x7	51.2	23.3	21.1	95.5	42.6	44.6	12.4	15.1
23	x2;x3;x4;x5;x6;x8	90.7	232.7	194.7	177.1	170.0	86.9	188.4	193.2
24	x2;x3;x4;x5;x7;x8	50.8	24.3	21.4	98.5	44.4	43.8	12.6	15.3
25	x2;x3;x4;x6;x7;x8	477.1	1,259.4	678.9	928.0	584.8	440.4	520.7	750.4
26	x2;x3;x5;x6;x7;x8	462.5	1,139.8	688.0	968.4	541.4	414.0	504.4	832.3
27	x2;x4;x5;x6;x7;x8	459.1	1,136.8	658.2	896.2	532.2	395.7	531.0	845.2
28	x3;x4;x5;x6;x7;x8	472.2	1,072.1	618.8	782.9	463.3	387.7	516.3	836.4

[1]Cases. [2]Traits: main stem height (x1), plant stem height (x2), plant stem length (x3), plant ear length (x4), plant peduncle length (x5), number of stems plant^{-1} (x6), number of nodes stem^{-1} (x7) and number of nodes plant^{-1} (x8). [3]Trials conducted in five and three sowing times with the cultivars BRS Progresso (T1, T2, T3, T4 and T5) and Temprano (T6, T7 and T8), respectively. *Cases disregarded in the present study, due to the high degree of multicollinearity.

Of these 28 cases, six cases were discarded in which the estimates of the degree of multicollinearity were extremely severe ($8.28\times10^{15} \leq CN \leq 3.36\times10^{17}$). Thus, 176 cases were considered (8 trials × 22 cases trial^{-1}).

The sample size was determined for estimating the indicators of the degree of multicollinearity - CN, DET and VIF - for each of the 176 cases. For this, in each case, 197 sample sizes were planned. The first planned sample size was composed of observations of 20 plants. The other planned sample sizes were obtained with the increment of five plants, up to the last size, containing 1,000 plants. Thus, in each case, the sample sizes of 20, 25, 30, ..., 1,000 plants were planned. Then, for each planned sample size, 2,000 resampling procedures with replacement were performed, and CN, DET and VIF were estimated in each one. After that, the mean degree of multicollinearity of each indicator in each planned sample size was calculated.

Finally, three models were fitted: modified maximum curvature method (MMCM), segmented linear model with plateau response (LMPR) and segmented quadratic model with plateau response (QMPR). In these three models, the mean of the indicator (CN, DET or VIF) (dependent variable, Y_i) was fitted as a function to the planned sample sizes (independent variable, X_i). For each case, indicator and model (176 × 3 × 3 = 1,584 situations), were determined the sample size (n), the multicollinearity degree obtained in the fitting corresponding to n ($CN_{(n)}$, $DET_{(n)}$ and $VIF_{(n)}$) and the adjusted coefficient of determination (R^2_a).

Coefficients a and b for MMCM were determined by the expression of Equation 1:

$$Y_i = a/X_i^b + \varepsilon_i \quad (1)$$

where: X_i is the independent variable, that is, the planned sample sizes (20, 25, 30, ..., 1,000 plants), and Y_i is the dependent variable referring to the value (mean of 2,000 estimates) of each indicator of the degree of multicollinearity. The sample size (n) was determined according to Equation 2 (MEIER; LESSMAN, 1971) and the estimate of the multicollinearity corresponding to n according to Equation 3, where *a* and *b* are the model parameters.

$$n = \left[\frac{a^2 b^2 (2b+1)}{(b+2)}\right]^{1/(2b+2)} \quad (2)$$

$$Y_{(n)} = a/n^b \quad (3)$$

Regarding the functions with plateau response, Equation 4 was considered for LMPR and Equation 5 was considered for QMPR:

$$Y_i \begin{cases} a + bX_i + \varepsilon_i & \text{if } X_i \leq n \\ P + \varepsilon_i & \text{if } X_i > n \end{cases} \quad (4)$$

$$Y_i \begin{cases} a + bX_i + cX_i^2 + \varepsilon_i & \text{if } X_i \leq n \\ P + \varepsilon_i & \text{if } X_i > n \end{cases} \quad (5)$$

where: X_i is the independent variable, that is, the planned sample sizes (20, 25, 30, ..., 1,000 plants); Y_i is the dependent variable referring to the value (mean of 2,000 estimates) of the degree of multicollinearity of each indicator; *a*, *b* and *c* are the parameters of the models; ε_i is the error associated with the i-th observation; P is the plateau; and n is the estimate of the sample size and the point of union between the two functions.

The n parameter was determined considering the union between the two lines for LMPR and QMPR according to Equation 6. For the estimation of the degree of

multicollinearity ($Y_{(n)}$), the estimate of P was considered for LMPR and Equation 7 was considered for QMPR, where $\hat{a}$, $\hat{b}$ and $\hat{c}$ are the estimates of the model parameters.

$$n = -\hat{b}/(2\times\hat{c}) \quad (6)$$

$$Y_{(n)} = \hat{a} - \hat{b^2}/(4\times\hat{c}) \quad (7)$$

For each trial, indicator and model, were calculated the minimum, maximum and mean values of the sample size (n), the estimation of the degree of multicollinearity obtained in the fitting of the model for n ($Y_{(n)} = CN_{(n)}$ or $DET_{(n)}$ or $VIF_{(n)}$) and the adjusted coefficient of determination (R^2_a), among the 22 cases. The means of R^2_a for each indicator were taken into account for choosing the model to be used in the inference of n. After defining the model, the mean estimates of the sample size of each trial were compared through a Scott-Knott means comparison test, at 5% significance level and the means of the sample size among the indicators of the same model, respectively, were compared at 5% significance level by the Student's t-test for independent samples. The fits by QMPR of the degree of multicollinearity of the three indicators, as well as the cases of lowest and highest degree of multicollinearity, were graphically presented. Statistical analyses were carried out in R software (R TEAM CORE, 2019).

RESULTS AND DISCUSSION

The existence of a severe degree of multicollinearity (CN > 1,000) (MONTGOMERY; PECK; VINNING, 2012) was verified in the master sample in trials when considering the eight morphological traits of rye, with the values of condition number (CN) higher than 1.5×10^{16} (Table 1). Similarly, severity was verified for all cases when combining seven traits. For 28 cases obtained by the combination of six traits, trials with weak (CN ≤ 100), moderate to strong (100 < CN ≤ 1,000) and severe multicollinearity (CN > 1,000) (MONTGOMERY; PECK; VINNING, 2012) were observed for the master sample.

As cases 1, 2, 3, 16, 17 and 18 showed a severe degree of multicollinearity ($8.28\times10^{15} \leq CN \leq 3.36\times10^{17}$) and due to the impossibility of resampling, these cases were disregarded in the present study. Severe multicollinearity causes the data matrix to be poorly conditioned and consequently a source of computational error, leading to signal errors and parameters of different magnitudes (MONTGOMERY; PECK; VINNING, 2012). Thus, among the 176 cases (8 trials × 22 cases $trial^{-1}$), 26.14% showed estimates of weak (CN ≤ 100), 65.91% showed estimates of moderate to strong (100 < CN ≤1,000) and 7.95% showed estimates of severe multicollinearity (CN > 1,000) (MONTGOMERY; PECK; VINNING, 2012).

The degree of multicollinearity obtained by the CN, correlation matrix determinant (DET) and variance inflation factor (VIF) of the master sample, in the 22 cases and in each trial, were presented only in a summarized way in Table 2, which showed: 12.36 ≤ CN ≤ 1,401.96; 0.000019 ≤ DET ≤ 0.165307; and 3.25 ≤ VIF ≤ 196.27, with greater variability of multicollinearity estimates among the cases observed for the indicator DET (coefficient of variation - $CV_{DET} \geq 163.46\%$). The estimates obtained by the other two indicators also showed high variability, but of lower magnitudes ($63.99\% \leq CV_{CN} \leq 87.74\%$ and $63.51\% \leq CV_{VIF} \leq 89.01\%$).

This variability of multicollinearity estimates was due to the cases, which are formed by the combination of eight traits taken six by six. A study with rye characteristics to assess the relationship between grain yield and yield and morphological components reported variability in the estimates of multicollinearity (1.37 ≤ VIF ≤ 452) and traits were removed from the regression model with VIF > 10 (NOURAEIN, 2019). In a trial with wheat crop, it was not necessary to eliminate traits because VIF was lower than 1.46 (JANMOHAMMADI; SABAGHNIA; NOURAEIN, 2014). However, in maize, the VIF estimate was higher than 195.58, using all observations or mean values per plot in the diagnosis of multicollinearity (OLIVOTO et al., 2017b).

No studies with rye crop in which diagnoses were made by CN or DET were found. In other crops, low degree of multicollinearity was observed in sunflower traits (CN = 9.64) (FOLLMANN et al., 2019) and severe multicollinearity was observed in morphological traits of showy rattlepod (CN = 1,113.08) (TOEBE et al., 2017a) and maize hybrids (CN > 1,000) (TOEBE; CARGNELUTTI FILHO, 2013; OLIVOTO et al., 2017b; TOEBE et al., 2017b). The DET was used for the diagnosis in maize traits using all observations ($DET = 3.02\times10^{-6}$) and mean values of plots ($DET = 1.26\times10^{-7}$) (OLIVOTO et al., 2017b). In cherry tomatoes, DET values between 0.00002 and 0.02500 were obtained in a study on the impact of sample size on the degree of multicollinearity (SARI et al., 2018). These studies demonstrate that high estimates of multicollinearity can be obtained, regardless of the indicator.

Both in this study and in the other studies presented above, there was variation in the estimate or occurrence of absence or high multicollinearity. As it can be defined as the relationship between traits (MONTGOMERY; PECK; VINNING, 2012), the different levels of multicollinearity are due to the traits and their interrelationships. Thus, the researcher should conduct a survey in the literature and choose sets of traits that capture as much as possible the variability of the phenomenon under study and that have the lowest degree of multicollinearity. Therefore, it is important to know the traits that, when combined, can cause collinearity, thus preventing them from being evaluated and then eliminated later when conducting multivariate analysis. Thus, in order to avoid evaluating traits that may cause problems due to multicollinearity, the researcher should choose traits from any of the cases with CN ≤ 100 (Table 1).

Table 2. Minimum (Min), maximum (Max), mean, standard deviation (SD) and coefficient of variation (CV) of the estimates of the degree of multicollinearity obtained by three indicators (CN, DET and VIF), determined from the master sample (n master), in 22 cases and in eight uniformity trials with rye crop (*Secale cereale* L.), conducted in the 2016 season, Santa Maria, RS, Brazil.

Trial	Cultivar[1]	Time[2]	n master	Cases	Min	Max	Mean	SD	CV (%)
					---------- Condition number (CN) ----------				
T1	BRSP	1	100	22	50.78	603.52	290.97	199.15	68.44
T2	BRSP	2	100	22	23.27	1,401.96	609.79	535.03	87.74
T3	BRSP	3	100	22	21.06	885.13	421.36	298.49	70.84
T4	BRSP	4	90	22	95.52	1,151.17	540.53	390.11	72.17
T5	BRSP	5	100	22	42.57	675.49	382.54	246.27	64.38
T6	TEMP	1	100	22	43.79	538.18	284.00	181.74	63.99
T7	TEMP	2	100	22	12.36	674.67	322.63	227.49	70.51
T8	TEMP	3	90	22	15.09	1,029.16	509.45	391.24	76.80
Mean	-	-	-	-	38.0547	869.9095	420.1565	308.6883	71.86
					---------- Determinant (DET) ----------				
T1	BRSP	1	100	22	0.000107	0.052185	0.010212	0.016692	163.46
T2	BRSP	2	100	22	0.000066	0.095446	0.018168	0.031870	175.42
T3	BRSP	3	100	22	0.000055	0.107532	0.018739	0.034722	185.29
T4	BRSP	4	90	22	0.000019	0.025850	0.004304	0.007540	175.16
T5	BRSP	5	100	22	0.000037	0.037124	0.006399	0.011249	175.78
T6	TEMP	1	100	22	0.000115	0.040365	0.007405	0.012373	167.08
T7	TEMP	2	100	22	0.000116	0.165307	0.030133	0.054731	181.63
T8	TEMP	3	90	22	0.000043	0.162844	0.027501	0.049901	181.45
Mean	-	-	-	-	0.000700	0.085832	0.015358	0.027385	175.66
					---------- Variance inflation factor (VIF) ----------				
T1	BRSP	1	100	22	11.28	105.32	60.43	42.31	70.01
T2	BRSP	2	100	22	5.37	235.42	120.30	107.08	89.01
T3	BRSP	3	100	22	4.71	151.84	86.78	63.84	73.56
T4	BRSP	4	90	22	21.23	196.27	108.36	80.49	74.28
T5	BRSP	5	100	22	8.79	107.63	67.80	43.27	63.83
T6	TEMP	1	100	22	8.72	81.90	51.79	32.89	63.51
T7	TEMP	2	100	22	3.25	126.35	71.81	52.68	73.37
T8	TEMP	3	90	22	3.80	194.94	106.20	82.51	77.70
Mean	-	-	-	-	8.3934	149.9566	84.1830	63.1346	73.16

[1]Rye cultivars: BRS Progresso (BRSP) and Temprano (TEMP). [2]Sowing time 1, 2, 3, 4 and 5 on 5/3/2016, 05/25/2016, 6/7/2016, 6/22/2016, and 7/4/2016, respectively.

The sample sizes (n) in each case and trial were obtained by fitting the degree of multicollinearity according to the sample size using three models, and the means for each trial are presented in Table 3. The worst fits for the three indicators were verified in the modified maximum curvature method (MMCM). For this model, mean values of adjusted coefficients of determination (R^2_a) of each trial and in each indicator differed at 5% probability level by the Student's t-test for independent samples (Table 4), when compared with the other two models: $0.56 \leq R^2_a \leq 0.68$, $0.58 \leq R^2_a \leq 0.74$ and $0.50 \leq R^2_a \leq 0.64$ for the indicators CN, DET and VIF, respectively.

Table 3. Means of sample size (n), estimate of the degree of multicollinearity and adjusted coefficient of determination (R^2_a), obtained with the fit of three models of the condition number (CN), determinant (DET) and variance inflation factor (VIF), in rye uniformity trials (*Secale cereale* L.).

Trial[1]	Cases[2]	Sample size (n)			Estimate			R^2_a		
		CN	DET	VIF	CN	DET	VIF	CN	DET	VIF
---------- Modified maximum curvature method (MMCM) ----------										
T1	22	3,908	0	1,620	247	0.00268	52	0.66	0.70	0.55
T2	22	12,252	0	1,445	477	0.00479	107	0.68	0.70	0.62
T3	22	7,937	0	205	335	0.00479	94	0.62	0.71	0.57
T4	22	15,348	0	1,401	2,240	0.00121	99	0.58	0.66	0.52
T5	22	7,760	0	5,717	594	0.00281	63	0.56	0.58	0.50
T6	22	1,187	0	85	273	0.00200	62	0.64	0.69	0.58
T7	22	61,905	0	3,652	159	0.00728	55	0.66	0.74	0.64
T8	22	13,201	0	352	373	0.00767	110	0.61	0.70	0.59
Mean	-	15,437	0	1,810	587.53	0.00415	80.25	0.63	0.68	0.57
---------- Segmented Linear Model with Plateau Response (LMPR) ----------										
T1	22	105	128	65	301	0.01001	63	0.85	0.88	0.88
T2	22	104	123	75	635	0.01771	125	0.87	0.88	0.87
T3	22	79	127	67	436	0.01825	89	0.86	0.88	0.86
T4	22	89	114	58	556	0.00424	112	0.86	0.86	0.87
T5	22	82	200	61	396	0.00633	70	0.87	0.74	0.87
T6	22	96	121	70	292	0.00726	53	0.87	0.87	0.86
T7	22	86	145	77	341	0.02926	76	0.87	0.89	0.87
T8	22	73	124	71	519	0.02683	109	0.86	0.88	0.86
Mean	-	89	135	68	434.50	0.01499	87.13	0.87	0.86	0.87
---------- Segmented Quadratic Model with Plateau Response (QMPR) ----------										
T1	22	141a	174b	81b	300	0.01001	63	0.91	0.91	0.92
T2	22	140a	168b	97a	634	0.01773	125	0.91	0.91	0.91
T3	22	100b	173b	84b	436	0.01827	89	0.90	0.91	0.90
T4	22	112b	152b	70c	556	0.00425	112	0.90	0.90	0.91
T5	22	102b	236a	68c	396	0.00633	70	0.90	0.80	0.91
T6	22	124a	166b	87b	292	0.00726	53	0.91	0.91	0.90
T7	22	111b	199a	99a	340	0.02929	76	0.91	0.92	0.90
T8	22	97b	169b	91b	519	0.02686	109	0.90	0.91	0.90
Mean	-	116	180	85	434.13	0.01500	87.13	0.90	0.90	0.90

[1]Trials described in Table 2. [2]Traits for each case described in Table 1.

The segmented linear (LMPR) and quadratic (QMPR) models with plateau response showed estimates of $R^2_a \geq 0.74$ and very similar to each other, but with superiority of the means of R^2_a for QMPR in the fit of CN and VIF, at 5% probability level (Table 4). For this model, the mean estimates of R^2_a between the trials for the CN, DET and VIF indicators were $0.90 \leq R^2_a \leq 0.91$, $0.80 \leq R^2_a \leq 0.92$ and $0.90 \leq R^2_a \leq 0.92$, respectively. Models are considered to be of good fit when R^2_a values are greater than 0.80.

Due to the superiority obtained by QMPR in fitting the degree of multicollinearity as a function of sample size, this model was selected to determine n for the CN, DET and VIF indicators. For each indicator, the fits by QMPR of the trials that had the lowest and highest degree of multicollinearity among the data of the master sample were presented in graphs (Figure 1).

Table 4. Comparison of means of adjusted coefficient of determination between three models for each indicator and means of the sample size between the indicators for the segmented quadratic model with plateau response by Student's t-test for independent samples.

	Adjusted coefficient of determination (R^2_a)								
Indicator[1]	MMCM[2]	LMPR	p-value	MMCM	QMPR	p-value	LMPR	QMPR	p-value
CN	0.63	0.87	<0.001	0.63	0.90	<0.001	0.87	0.90	<0.001
DET	0.68	0.86	<0.001	0.68	0.90	<0.001	0.86	0.90	0.126
VIF	0.57	0.87	<0.001	0.57	0.90	<0.001	0.87	0.90	<0.001
	Segmented Quadratic Model with Plateau Response (QMPR)								
Sample size	CN	DET	p-value	CN	VIF	p-value	DET	VIF	p-value
	116	180	<0.001	116	85	<0.001	180	85	<0.001

[1]CN = condition number; DET = correlation matrix determinant; VIF = variance inflation factor. [2]MMCM = modified maximum curvature method; LMPR = segmented linear model with plateau response; QMPR = segmented quadratic model with plateau response.

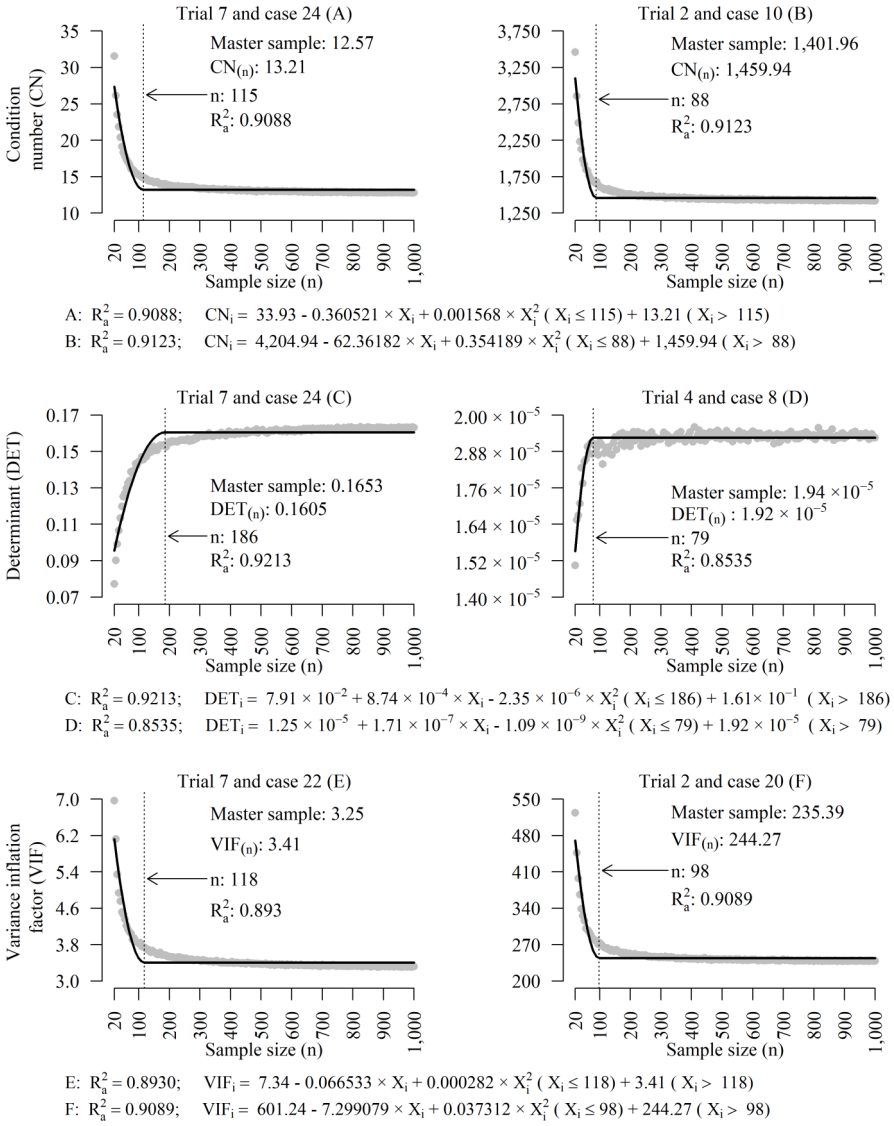

Figure 1. Sample size (n) estimated by the segmented quadratic model with plateau response (QMPR) for the indicators condition number (CN), correlation matrix determinant (DET) and variance inflation factor (VIF), and the respective estimated multicollinearity for each indicator ($CN_{(n)}$, $DET_{(n)}$ and $VIF_{(n)}$) in morphological traits evaluated in eight uniformity trials of rye (*Secale cereale* L.). Trial and case with the lowest [A, C and E] and highest [B, D and F] estimate of CN, DET and VIF, respectively, in the master sample.

The n necessary for the diagnosis of the degree of multicollinearity between morphological traits of rye, obtained through QMPR, varied among the 176 cases (8 trials × 22 cases trial^{-1}). The middle n among the trials and cases was 116, with the variation in the mean values of n within each trial of 97 ≤ n ≤ 141 for CN. The estimation of the degree of multicollinearity by the DET indicator requires a larger sample size (n) or number of plants, with a mean value of 180 and between the trials, with means 36 ≤ n ≤ 859. Among the three indicators, the detection of the multicollinearity degree by VIF requires the lowest n, with an overall mean of 85 plants and means of 68 ≤ n ≤ 99 for the trial of highest and lowest estimate of n. Due to the significant differences in n means between indicators, it can be affirmed that there is variability among the estimates by CN, DET and VIF indicators in morphological traits of rye. This demonstrates the need to contemplate in the experimental planning also the indicator to be used in the diagnosis of multicollinearity. Given the significant difference and aiming at greater precision, larger size of n should be used, with n = 180 plants (mean value of plants obtained by DET).

It can also be observed that there is variability in the sample size estimates to detect the degree of multicollinearity among the trials (T1 to T8). Thus, sowing time has an effect on the average estimates of n in the same cultivar (T1 to T5 for the cultivar BRS Progresso and T6 to T8 for the cultivar Temprano) and among the cultivars. When comparing the estimates of n between sowing times, there was also no standard behavior of the highest mean of n from one indicator to another. Considering the CN indicator, the highest means were observed in the trials corresponding to the first sowing time in both cultivars (T1 and T6) and the second sowing time for the cultivar Temprano; whereas for DET, the highest means were observed in the trials corresponding to the fifth sowing time for BRS Progresso (T5) and second sowing time for Temprano (T7); second sowing time in both cultivars (T2 and T7) for VIF. Effects of sowing time and rye cultivar were also observed in studies to determine the sample size to estimate the mean value of morphological traits and in flowering stage (BANDEIRA et al., 2018a; 2018b).

No studies with rye crop in which the sample size study was performed for the diagnosis of multicollinearity were found. Some inferences have been made in studies with maize and cherry tomato, indicating that insufficient sample sizes could incorrectly estimate the degree of multicollinearity (OLIVOTO et al., 2017a; SARI et al., 2018).

Olivoto et al. (2017a) point out that problems caused by multicollinearity can be mitigation by using all observations to generate the correlation matrix, instead of using the mean values. The authors used data considering all observations or grouped data for the mean and found that the lower the number of observations (use of means), the greater the inaccuracy in the estimates. Sari et al. (2018) found the need for sample sizes greater than 45 plants to estimate multicollinearity by the DET indicator, with a 5% probability of error using the *bootstrap* methodology with a 95% confidence interval, and that when using sample size greater than 135 plants there would be no interference of the sample in the diagnosis of the degree of multicollinearity.

However, the present study found, for morphological traits of rye, the need for sample size of at least 180 plants, a value higher than that reported by Sari et al. (2018) in cherry tomato traits. This difference may be associated with the species, evaluated traits or methodology used in the determination of sample size. However, the results of this study point to the need for a sample size greater than 135 plants in rye. Therefore, further investigations on sample size for the diagnosis of multicollinearity in the most diverse agricultural crops should be carried out to check for possible variability of n.

Significant differences between the means of sample size, at 5% significance level, were verified by the Student's t-test for independent samples, in the comparisons of CN × DET, CN × VIF and DET × VIF (Table 4), considering the values of 176 sample sizes (8 trials × 22 trial^{-1}) for each indicator. These results confirm that higher values of n or number of plants are necessary when using the DET indicator, followed by CN and VIF, for the diagnosis of the degree of multicollinearity in correlation matrices of rye morphological traits.

In this study, it was found that it is necessary to use different sample size when diagnosing multicollinearity by the indicators condition number, correlation matrix determinant and variance inflation factor in morphological traits of rye. Aiming at greater precision, larger sample sizes should be prioritized, adopting an average size determined for the correlation matrix determinant indicator (n = 180 plants). As a method to determine the sample size, it is not recommended to use the modified maximum curvature method, but rather the segmented quadratic model with plateau response. Other models should be investigated for the possibility of use in determining the sample size.

CONCLUSIONS

There is variability in sample size between the indicators condition number (CN), correlation matrix determinant (DET) and variance inflation factor (VIF) for the diagnosis of the degree of multicollinearity in morphological traits of rye, with increase in the following order: VIF, CN and DET, which require at least 85, 116 and 180 plants, respectively. If there is interest in greater precision, a larger sample size should be prioritized, with the adoption of sample size obtained for the DET indicator.

ACKNOWLEDGMENTS

To the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq - National Council for Scientific and Technological Development; Processes 401045/2016-1, 304652/2017-2, and 146258/2019-3), *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES - Coordination for the Improvement of Higher Education Personnel), and *Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul* (FAPERGS - Rio Grande do Sul State Research Support Foundation) for the scholarships granted.

REFERENCES

ABOU CHEHADE, L. et al. Rye (*Secale cereale* L.) and squarrose clover (*Trifolium squarrosum* L.) cover crops can increase their allelopathic potential for weed control when used mixed as dead mulch. **Italian Journal of Agronomy**, 16: 1–11, 2021.

ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, 22: 711-728, 2013.

ALVES, B. M.; CARGNELUTTI FILHO, A.; BURIN, C. Multicollinearity in canonical correlation analysis in maize. **Genetics and Molecular Research**, 16: 1–14, 2017.

BAIER, A. C. **Centeio**. Passo Fundo, RS: EMBRAPA-CNPT, 1994. 29 p. (Documentos, 15).

BANDEIRA, C. T. et al. Sample size to estimate the mean of morphological traits of rye cultivars in sowing dates and evaluation times. **Semina: Ciências Agrárias**, 39: 521-532, 2018a.

BANDEIRA, C. T. et al. Sample sufficiency for estimation of the mean of rye traits at flowering stage. **Journal of Agricultural Science**, 10: 178-186, 2018b.

BASCHE, A. D. et al. Soil water improvements with the long-term use of a winter rye cover crop. **Agricultural Water Management**, 172: 40–50, 2016.

FIELD, A. **Descobrindo a estatística utilizando o SPSS**. 2 ed. Porto Alegre, RS: Artmed, 2009. 688 p.

FOLLMANN, D. N. et al. Correlations and path analysis in sunflower grown at lower elevations. **Journal of Agricultural Science**, 11: 445-453, 2019.

GUJARATI, D. N.; PORTER, D. C. **Econometria básica**. 5 ed. Porto Alegre, RS: AMGH Editora Ltda, 2011. 920 p.

JANMOHAMMADI, M.; SABAGHNIA, N.; NOURAEIN, M. Path Analysis of Grain Yield and Yield Components and Some Agronomic Traits in Bread Wheat. **Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis**, 62: 945-952, 2014.

LAIDIG, F. et al. Breeding progress, variation, and correlation of grain and quality traits in winter rye hybrid and population varieties and national on-farm progress in Germany over 26 years. **Theoretical and Applied Genetics**, 5: 981-998, 2017.

MEIER, V. D.; LESSMAN, K. J. Estimation of optimum field plot shape and size for testing yield in *Crambe abyssinica* Hochst. **Crop Science**, 11: 648-650, 1971.

MONTGOMERY, D. C.; PECK, E. A. VINNING, G. G. **Introduction to linear regression analysis**. New York: John Wiley and Sons, 2012. 672 p.

MORRISON, L. A. Cereals: Domestication of the Cereal Grains. **Encyclopedia of Food Grains**. 1: 86-98, 2016.

NOURAEIN, M. Elucidating seed yield and components in rye (*Secale cereale* L.) using path and correlation analyses. **Genetic Resources and Crop Evolution**, 66: 1533-1542, 2019.

OLIVOTO, T. et al. Optimal sample size and data arrangement method in estimating correlation matrices with lesser collinearity: A statistical focus in maize breeding. **African Journal of Agricultural Research**, 12: 93-103, 2017a.

OLIVOTO, T. et al. Multicollinearity in path analysis: A simple method to reduce its effects. **Agronomy Journal**, 109: 131-142, 2017b.

PAULINO, V. T.; CARVALHO, D. D. Pastagens de inverno. **Revista Científica Eletrônica de Agronomia**, 3: 1-6, 2004.

R TEAM CORE. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, 2019. Disponível em: <https://www.r-project.org/>. Acesso em: 12 dez. 2019.

SANTOS, H. G. et al. **Brazilian Soil Classification System**. 5 ed. Brasília, DF: EMBRAPA , 2018. 469 p.

SAPIRSTEIN, H. D.; BUSHUK, W. Rye Grain: Its Genetics, Production, and Utilization. **Encyclopedia of Food Grains**, 1: 159-167, 2016.

SARI, B. G. et al. Interference of sample size on multicollinearity diagnosis in path analysis. **Pesquisa**

Agropecuária Brasileira, 53: 769-773, 2018.

TOEBE, M.; CARGNELUTTI FILHO, A. Não normalidade multivariada e multicolinearidade na análise de trilha em milho. **Pesquisa Agropecuária Brasileira**, 48: 466-477, 2013.

TOEBE, M. et al. Dimensionamento amostral e associação linear entre caracteres de *Crotalaria spectabilis*. **Bragantia**, 76: 45-53, 2017a.

TOEBE, M. et al. Direct effects on scenarios and types of path analyses in corn hybrids. **Genetics and Molecular Research**, 16: 1-15, 2017b.