PAB
Pesquisa Agropecuária Brasileira

ISSN 1678-3921

Journal homepage: www.embrapa.br/pab

For manuscript submission and journal contents, access: www.scielo.br/pab

Statistics/ Original Article

Sample size for principal component analysis in corn

Abstract – The objective of this work was to determine the number of plants required to estimate the eigenvalues of the principal components analysis in corn (*Zea mays*) traits. Twelve traits were measured in 361, 373, and 416 plants of single-, three-way, and double-cross hybrids, respectively, in the 2008/2009 crop year; and in 1,777, 1,693, and 1,720 plants of single-, three-way, and double-cross hybrids, respectively, in the 2009/2010 crop year (six cases), totaling 6,340 plants. Principal component analysis was performed for the six cases. Sample size (number of plants) for the eigenvalue estimations of the principal components was determined by resampling with replacement and application of the model linear response and plateau model. The measurement of 267 plants is sufficient to estimate the eigenvalues of the principal components in corn traits.

Index terms: *Zea mays*, model linear response and plateau model, multivariate analysis, resampling.

Tamanho de amostra para análise de componentes principais em milho

Resumo – O objetivo deste trabalho foi determinar o número de plantas necessário para estimar os autovalores dos componentes principais em caracteres de milho (*Zea mays*). Doze caracteres foram mensurados em 361, 373 e 416 plantas de híbridos simples, triplo e duplo, respectivamente, no ano agrícola de 2008/2009; e em 1.777, 1.693 e 1.720 plantas de híbridos simples, triplo e duplo, respectivamente, no ano agrícola de 2009/2010 (seis casos), no total de 6.340 plantas. As análises de componentes principais foram realizadas para os seis casos. Determinou-se o tamanho de amostra (número de plantas) para a estimação dos autovalores dos componentes principais, por reamostragem com reposição e com aplicação do modelo linear de resposta com platô. A mensuração de 267 plantas é suficiente para estimar os autovalores dos componentes principais em caracteres de milho.

Termos para indexação: *Zea mays*, modelo linear de resposta com platô, análise multivariada, reamostragem.

Introduction

Corn crop researches have been intensively carried out with a high number of variables for the evaluation and discrimination of new genotypes, as well as the identification of optimal cultivation conditions and limiting factors for productivity. Such researches are conducted in laboratories, greenhouses, experimental areas and trial network, and they are associated with experimental errors that − if not properly controlled or circumvented − may affect the power available to reject a null hypothesis (Dochtermann & Jenkins, 2011). In this

Alberto Cargnelutti Filho[1 ✉] and Marcos Toebe[2],

[1] Universidade Federal de Santa Maria, Departamento de Fitotecnia, Avenida Roraima, nº 1.000, Camobi, CEP 97105-900
Santa Maria, RS, Brazil.
E-mail: alberto.cargnelutti.filho@gmail.com

[2] Universidade Federal de Santa Maria, Departamento de Ciências Agronômicas e Ambientais, Campus Frederico Westphalen, Linha 7 de Setembro, s/nº, BR-386, Km 40, CEP 98400-000 Frederico Westphalen, RS, Brazil.
E-mail: m.toebe@gmail.com

✉ Corresponding author

Received
April 28, 2021

Accepted
October 22, 2021

How to cite
CARGNELUTTI FILHO, A.; TOEBE, M. Sample size for principal component analysis in corn. **Pesquisa Agropecuária Brasileira**, v.56, e02510, 2021. DOI: https://doi.org/10.1590/S1678-3921.pab2021.v56.02510.

sense, the knowledge of the appropriate sample size and the corresponding precision are important to define experimental protocols for the crop. Studies using resampling techniques to define the appropriate sample size for estimating the mean and coefficient of variation (Toebe et al., 2014), correlations (Cargnelutti Filho et al., 2010; Toebe et al., 2015; Olivoto et al., 2017, 2018), path analysis (Toebe et al., 2017) and multiple regression (Cargnelutti Filho & Toebe, 2020) have been carried out on corn cultivation. Such studies have shown variability between techniques, variables, hybrids, scenarios, and precision levels.

The increase of data processing capacity and the availability of softwares and statistical packages have led researchers from multiple areas to apply more complex data analysis techniques in the evaluation of their experiments, especially when a high number of variables are evaluated. To facilitate the interpretation of these data, principal component analysis (PCA) can be applied, whose main purpose is to reduce the dimensionality of multivariate data and to facilitate the interpretation of results by generating new variables (components) (Lattin et al., 2011). In PCA, the amount of information is maximized in the first components, especially in the case of variables with a high redundancy index (Lattin et al., 2011), allowing of the inference on the phenomena under study (Ferreira, 2018).

For PCA and other correlate multivariate methods, some studies on sample size and general recommendations have been carried out (Stauffer et al., 1985; Osborne & Costello, 2004; Ramachandran & Aschheim, 2005; Kocovsky et al., 2009; Dochtermann & Jenkins, 2011; Shaukat et al., 2016; Björklund, 2019; Gañan-Cardenas & Correa-Morales, 2021). Starting from a small sample (n = 55 observations) and a high number of variables (p = 22), Shaukat et al. (2016) simulated four sample sizes (n = 20, 30, 40, and 50) and justified the use of the database by cost of water quality analysis. The authors concluded that a sample size of 40 or 50 is sufficient for ecological and environmental studies to recover the first few components. According to Björklund (2019), the robustness of the principal components increases with increasing sample size, but not with the number of traits. Still, in a study on the inferential process, Gañan-Cardenas & Correa-Morales (2021) empirically indicated the use of a subject to item ratio of 10:1 and 20:1, for PCA estimate from the covariance and correlation matrix, respectively. According to Kocovsky et al. (2009), minimum sample size recommendations are rarely accompanied by empirical support. In several studies on corn cultivation, PCA was applied with the following aims: to characterize hybrids for water shortage (Guimarães et al., 2014); to characterize grain yield, and other variables, in different corn hybrids grown under heat and drought stress (Ali et al., 2015); to predict flowering time, yield, and kernel dimensions by analyzing aerial images (Wu et al., 2019); and to characterize corn populations (Belalia et al., 2019). However, we could not find in the literature any study indicating the optimal sample size for the application of PCA in real data for corn crop.

The objective of this work was to determine the sample size required to estimate the eigenvalues of the principal components analysis of corn traits.

Materials and Methods

Two experiments were carried out with corn, in an experimental area located at 29º 42' S, 53º 49' W, 95 m altitude, in Santa Maria, Rio Grande do Sul state, Brazil. The first experiment was conducted in the 2008/2009 crop year. The second experiment was conducted in the 2009/2010 crop year. According to the Köppen-Geiger classification, the climate of the region is Cfa, subtropical humid (Alvares et al., 2013). The soil is classified as Argissolo Vermelho distrófico, according to the Brazilian soil classification system (Santos et al., 2018), that corresponds to Ultisol classification (Soil Survey Staff, 1999).

In the first experiment, sowing was performed on December 26, 2008. Four plots were sown with the single-cross hybrid P32R21, four with the three-way cross hybrid DKB566, and four with the double-cross hybrid DKB747. In the second experiment, sowing was performed on October 26, 2009. Sixteen plots were sown with the single-cross hybrid 30F53, sixteen with the three-way cross hybrid DKB566, and sixteen with the double-cross hybrid DKB747.

Each plot consisted of four 6.0 m rows, 0.8 m apart, with density adjusted to five plants per row meter, representing the density of 62,500 plants per hectare. Thus, each plot consisted of 120 plants, totaling 1,440 plants in the first experiment (3 hybrids × 4 plots/hybrid × 120 plants/plot), and 5,760 plants in the second experiment (3 hybrids × 16 plots/hybrid × 120 plants/plot). In each crop year, plots of the single-, three-way, and double-cross hybrids were randomized in the experimental area. In the two experiments, base fertilization was performed with 22.5 kg ha^{-1} N, 180 kg ha^{-1} P_2O_5, and 135 kg ha^{-1} K_2O, and the topdressing was 135 kg ha^{-1} N. The other cultural practices were performed according to the recommendations for corn cultivation (Fancelli & Dourado Neto, 2009).

In the first experiment, 361, 373, and 416 plants were assessed, for single-, three-way, and double-cross hybrids respectively. In the second experiment, 1,777, 1,693, and 1,720 plants were evaluated, for single-, three-way, and double-cross hybrids respectively. Only plants showing the twelve traits were evaluated, therefore, the final number of plants oscillated between plots and hybrids. Thus, the following traits were measured for 6,340 plants, as follows: plant height at harvest (PH, in cm); ear insertion height (EIH, in cm); ear weight (EW, in g); number of grain rows per ear (NR); ear length (EL, in cm); ear diameter (ED, in mm); cob weight (CW, in g); cob diameter (CD, in mm); hundred-grain weight (HGW, g); number of grains per ear (NGE); grain length (GL, in mm) calculated as the difference between the diameters of ear and cob divided by two; and grain yield (GY, in g per plant).

The principal component analysis was performed for each hybrid in each experiment (six cases), from the Pearson's linear correlation matrix between twelve traits (PH, EIH, EW, NR, EL, ED, CW, CD, HGW, NGE, GL, and GY). The correlation matrix was chosen because of the different trait measurement scales.

The sample size (n_o, number of plants) required to estimate the eigenvalues of principal component analysis was determined through resampling with replacement (Ferreira, 2009). For resampling, 986 sample sizes were planned, with an initial sample size of 15 plants (in this study considered as a reference, that is, minimum size required for principal component analysis). The other sizes were obtained in increments of one unit, until reaching 1,000 plants. Thus, sample sizes from 15 to 1,000 plants were planned.

For each planned sample size, 3,000 resamples with replacement were obtained. In each resample, the estimates of the eigenvalues of the twelve principal components (PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9, PC10, PC11, and PC12) were obtained. Therefore, for each sample size, 3,000 estimates of eigenvalues of the PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9, PC10, PC11, and PC12 were obtained, and the maximum, percentile 97.5% ($P_{97.5\%}$), mean, percentile 2.5% ($P_{2.5\%}$), and minimum were determined. The amplitude of 95% confidence interval was calculated by the expression: ACI = $P_{97.5\%} - P_{2.5\%}$. It should be interpreted that smaller is the ACI, the more accurate are the estimates of eigenvalues of the PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9, PC10, PC11, and PC12.

The sample size (n_o, number of plants) required to estimate the eigenvalues of the first four principal components (PC1, PC2, PC3, and PC4) was dimensioned because they explained at least 80% of the total data variation. Thus, for each hybrid in each experiment (six cases), the sample size (n_o, number of plants) was determined by adjusting the dependent variable [$ACI_{(n)}$] as a function of the independent variable (n, number of plants), by the model linear response with plateau (LRP) (Paranaíba et al., 2009).

For the LRP (Paranaíba et al., 2009), two segmented lines were adjusted, and the estimates of a, b and p parameters and the determination coefficient (r^2) were obtained. The first straight [$ACI_{(n)} = a + bn + \varepsilon$] was adjusted to the point corresponding to the optimal sample size (n_o), with nonnull slope (b). The second straight [$ACI_{(n)} = p + \varepsilon$] starts from n_o and has a zero slope, that is, it is a line parallel to the abscissa, in which p = plateau, that is, p corresponds to $ACIn_o$. The LRP model was

$$ACI_{(n)} = \begin{cases} a + bn + \varepsilon & \text{if } n \leq n_o \\ p + \varepsilon & \text{if } n > n_o \end{cases}$$

In the LRP model, the optimal sample size was determined by $n_o = (p - a) / b$ and the amplitude of the confidence interval in the optimal sample size by $ACIn_o = a + bn_o$; in which: the LRP model ACI is a dependent variable (amplitude of confidence interval of 95%); a is the intercept of the simple linear model of the segment previous to the plateau; b is the slope in this same segment; ε is the random error; p is the plateau; and n_o is the junction point of the two segments.

The percentile 97.5% ($P_{97.5\%}$), as well as the mean, percentile 2.5% ($P_{2.5\%}$), and amplitude of confidence interval of 95% (ACI) were plotted in graphs for better visual representation. The statistical analysis was performed using Microsoft Office Excel and the R software (R Core Team, 2021).

Results and Discussion

The eigenvalue estimates of the principal components (PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9, PC10, PC11, and PC12) were similar among hybrids and experiments (six cases) (Table 1). The eigenvalues ranged for the first four principal components, as follows: 6.55 to 7.30 (PC1); 1.46 to 1.69 (PC2); 1.11 to 1.41 (PC3); and 0.70 to 1.06 (PC4). In the mean of the six cases, these first four principal components respectively showed the eigenvalues 6.75, 1.58, 1.24, and 0.88, which explained the variances of 56.24%, 13.15%, 10.36%, and 7.33% and accumulated explained variances of 56.24%, 69.40%, 79.76%, and 87.09%. These results show the possibility of reducing the dimensionality of the set of 12 traits in four principal components that explain 87.09% of the total variation of the data.

Although there are different recommendations on the number of components to be maintained in the studies, Ferreira (2018) highlights the percentage of explained variance accumulated between 70% to 90% as sufficient. Therefore, the four most important components in the PCA were considered for the study of the sample size. The remaining eight components added together account for only 12.91% of the total variability and were disregarded (Table 1). For similar studies, Guimarães et al. (2014) used five variables to characterize corn hybrids for water shortage, and they found that in the vegetative stage, the first two components explained 99.52% of the total variance. In the flowering stage, the first two components explained 85.08% of the total variance, and in the grain swelling

Table 1. Estimates of the variance (eigenvalues) of twelve principal components (PC1, PC2, …, PC12), from twelve traits in corn hybrids (*Zea mays*) grown in two crop years.

Estimate	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Single-cross hybrid P32R21 (n=361 plants) in the 2008/2009 crop year												
Variance (eigenvalues)	6.58	1.46	1.11	0.95	0.85	0.38	0.30	0.22	0.14	0.01	0.00	0.00
Percentage of variance	54.86	12.16	9.26	7.88	7.05	3.19	2.47	1.85	1.18	0.10	0.00	0.00
Cumulative percentage of variance	54.86	67.02	76.29	84.17	91.22	94.41	96.87	98.72	99.90	100.00	100.00	100.00
Three-way cross hybrid DKB566 (n=373 plants) in the 2008/2009 crop year												
Variance (eigenvalues)	6.63	1.59	1.41	0.90	0.66	0.25	0.21	0.19	0.16	0.01	0.00	0.00
Percentage of variance	55.23	13.24	11.73	7.46	5.50	2.07	1.72	1.59	1.35	0.10	0.00	0.00
Cumulative percentage of variance	55.23	68.47	80.20	87.66	93.16	95.23	96.95	98.55	99.90	100.00	100.00	100.00
Double-cross hybrid DKB747 (n=416 plants) in the 2008/2009 crop year												
Variance (eigenvalues)	6.73	1.48	1.22	1.06	0.69	0.29	0.22	0.16	0.14	0.01	0.00	0.00
Percentage of variance	56.11	12.33	10.20	8.84	5.74	2.38	1.84	1.33	1.17	0.07	0.00	0.00
Cumulative percentage of variance	56.11	68.44	78.64	87.47	93.21	95.59	97.43	98.76	99.93	100.00	100.00	100.00
Single-cross hybrid 30F53 (n=1,777 plants) in the 2009/2010 crop year												
Variance (eigenvalues)	7.30	1.59	1.17	0.70	0.50	0.33	0.20	0.11	0.09	0.01	0.00	0.00
Percentage of variance	60.86	13.26	9.78	5.80	4.16	2.75	1.65	0.91	0.75	0.08	0.00	0.00
Cumulative percentage of variance	60.86	74.12	83.90	89.70	93.86	96.61	98.26	99.17	99.92	100.00	100.00	100.00
Three-way cross hybrid DKB566 (n=1,693 plants) in the 2009/2010 crop year												
Variance (eigenvalues)	6.70	1.66	1.26	0.79	0.60	0.39	0.24	0.21	0.14	0.01	0.00	0.00
Percentage of variance	55.79	13.85	10.48	6.57	4.96	3.21	2.03	1.77	1.21	0.12	0.00	0.00
Cumulative percentage of variance	55.79	69.64	80.12	86.70	91.66	94.87	96.90	98.67	99.88	100.00	100.00	100.00
Double-cross hybrid DKB747 (n=1,720 plants) in the 2009/2010 crop year												
Variance (eigenvalues)	6.55	1.69	1.29	0.89	0.65	0.36	0.22	0.20	0.13	0.01	0.00	0.00
Percentage of variance	54.60	14.08	10.72	7.42	5.44	3.02	1.87	1.67	1.07	0.11	0.00	0.00
Cumulative percentage of variance	54.60	68.68	79.40	86.83	92.26	95.28	97.15	98.82	99.89	100.00	100.00	100.00
Overall mean												
Variance (eigenvalues)	6.75	1.58	1.24	0.88	0.66	0.33	0.23	0.18	0.13	0.01	0.00	0.00
Percentage of variance	56.24	13.15	10.36	7.33	5.47	2.77	1.93	1.52	1.12	0.10	0.00	0.00
Cumulative percentage of variance	56.24	69.40	79.76	87.09	92.56	95.33	97.26	98.78	99.90	100.00	100.00	100.00

stage, the first two components explained 98.52% of the total variance. In twelve F1 single-cross corn hybrids and four crop growing seasons, Ali et al. (2015) evaluated sixteen variables and found that the first two components had variances of 43.5% and 24.4%, respectively. In Algerian corn populations, Belalia et al. (2019) evaluated fourteen agromorphological traits and eighteen simple sequence repeat markers and found that the first two components explained 43.04% and 12.40% of the total variation, respectively.

It was observed that with the increase of the number of plants, the mean of the 3,000 estimates of the eigenvalues of PC1, in the six cases, stabilizes and approaches the mean obtained with the 361 plants of the simple cross hybrid P32R21 (case 1, subject to item ratio = 30.08:1, that is, 361 plants/12 traits), 373 plants of the three-way cross hybrid DKB566 (case 2 subject to item ratio = 31.08:1), 416 plants of the double-cross hybrid DKB747 (case 3 subject to item ratio = 34.67:1), 1,777 plants of the single-cross hybrid 30F53 (case 4 subject to item ratio = 148.08:1), 1,693 plants of the three-way cross hybrid DKB566 (case 5 subject to item ratio = 141.08:1), and 1,720 plants of the double-cross hybrid DKB747 (case 6 subject to item ratio = 143.33:1). A similar pattern is observed in the other three principal components (PC2, PC3, and PC4). This suggests a possible bias in the estimates of the mean, in situations of sample insufficiency, which is more visible in the eigenvalues of the first two principal components (PC1 and PC2) that were overestimated, and in the fourth principal component (PC4) that was underestimated with insufficient sample size (Table 2, and Figures 1, 2, and 3). According to Ramachandran & Aschheim (2005), when the sample size increases, the errors become small and finally reach a constant value.

Therefore, the greater amplitudes of the confidence interval of the PC1, PC2, PC3, and PC4 eigenvalues, in the six cases, obtained from 15 plants in comparison with those obtained with 1,000 plants, show that with 15 plants the eigenvalue estimates are less accurate, which may result in inaccurate and biased PCA, when the sample is insufficient. Therefore, it can be inferred that PCA generated from a small number of plants should not be considered, and also that it is important and necessary to define the sample size for the generation of accurate PCA.

The amplitude of 95% confidence interval in ACI for the eigenvalue estimates of PC1, PC2, PC3, and PC4, in the six cases, gradually decreased with the increase in the number of plants (Figures 1, 2, and 3). This result is expected and indicates that the increase of the number of plants provides an improvement of the accuracy of estimates and, consequently, more reliable PCAs, as already verified for the estimation of the mean and coefficient of variation (Toebe et al., 2014), correlations (Cargnelutti Filho et al., 2010; Toebe et al., 2015; Olivoto et al., 2017), direct effects of path analysis (Toebe et al., 2017), and multiple regression (Cargnelutti Filho & Toebe, 2020) in corn. For PCA, Stauffer et al. (1985) observed that, as the sample size increased, the amplitude of the 95% confidence interval for the principal components decreased, indicating a precision gain. A sharp decrease in the ACI up to approximately 200 plants was observed for the eigenvalues of the 1st and 2nd principal components (PC1 and PC2), and a decrease of 300 plants for the eigenvalues of the 3rd and 4th principal components (PC3 and PC4), which suggests that such sample sizes would be sufficient (Figures 1, 2, and 3). Afterward, the decreases were smaller, which indicates that the work to measure more plants would result in insignificant benefits for the precision of the eigenvalue estimates of the principal components.

Based on model linear response with plateau, the sample size (n_o, number of plants) required to estimate the eigenvalues of the first four principal components (PC1, PC2, PC3, and PC4) was similar between hybrids and experiments (six cases). In the six cases, the sample sizes of the first two principal components (PC1 and PC2) were relatively smaller than those of the third and fourth principal components (PC3 and PC4) (Table 3). In the mean of the six cases, the sample sizes necessary to estimate the eigenvalues of PC1, PC2, PC3, and PC4 were, respectively, 234, 212, 297, and 323 plants. Although estimates of the eigenvalues of PC1, PC2, PC3, and PC4 from as many plants as possible should be aimed to guarantee reliable PCAs, it seems reasonable to estimate the eigenvalues based on 267 plants, which corresponds to the general mean of 24 sample sizes (six cases × four principal components). Above this number of plants, the gains for precision (decrease in ACI) are insignificant (Figures 1, 2, and 3).

In view of the results of the present study and the aforementioned inferences, it seems reasonable to accept that 267 plants (subject to item ratio = 22.25:1) are sufficient for PCA in corn. For other statistics and analyses applied in corn sample sizing studies, values similar to that of this study have also been recommended. In this sense, for hybrids of different genetic bases, Cargnelutti Filho et al. (2010) recommended 252 plants to estimate the correlation of 91 pairs of variables with ACI equal to 0.30 of Pearson's linear correlation coefficient. In a later study expanded by Toebe et al. (2015), on hybrids of different genetic bases and crops, the authors recommended 195 plants for the estimation of correlation coefficients with a maximum ACI of 0.35. Still studying correlations in corn, Olivoto et al. (2017) indicated that 210 plants are sufficient to estimate the r in the ACI < 0.30 and Olivoto et al. (2018) indicated 50 to 206 plants for estimating correlations with ACI of 0.30, depending on the magnitude of the correlations. To estimate the direct effect with maximum ACI of 0.35, Toebe et al. (2017) identified that 265 plants would be sufficient to estimate all the direct effects of explanatory variables on productivity, in crops, hybrids, and path analysis scenarios. Finally, Cargnelutti Filho & Toebe (2020) indicated that 260 plants are sufficient to adjust precise

Table 2. Maximum, percentile 97.5% ($P_{97.5\%}$), mean, percentile 2.5% ($P_{2.5\%}$), minimum, and amplitude of confidence interval of 95% (ACI = $P_{97.5\%}$ - $P_{2.5\%}$) for 3,000 estimates of eigenvalues of the first four principal components (PC1, PC2, PC3, and PC4), and estimates obtained from 3,000 resamples with replacement for n = 15 and 1,000 plants of corn hybrids (*Zea mays*) grown in two crop years.

PC	Maximum	$P_{97.5\%}$	Mean	$P_{2.5\%}$	Minimum	ACI	Maximum	$P_{97.5\%}$	Mean	$P_{2.5\%}$	Minimum	ACI
	n = 15 plants						n = 1000 plants					
Single-cross hybrid P32R21 (n=361 plants) in the 2008/2009 crop year												
PC1	9.53	8.50	6.81	5.05	3.87	3.45	6.96	6.82	6.59	6.37	6.15	0.45
PC2	3.57	2.87	1.98	1.27	0.79	1.61	1.65	1.58	1.47	1.36	1.29	0.22
PC3	2.36	1.87	1.28	0.76	0.52	1.11	1.29	1.21	1.12	1.03	0.95	0.18
PC4	1.63	1.30	0.84	0.45	0.24	0.85	1.08	1.02	0.95	0.87	0.80	0.15
Three-way cross hybrid DKB566 (n=373 plants) in the 2008/2009 crop year												
PC1	9.73	8.47	6.88	5.22	4.10	3.25	7.02	6.83	6.63	6.43	6.26	0.41
PC2	3.93	2.92	2.07	1.38	0.95	1.54	1.81	1.71	1.60	1.50	1.42	0.21
PC3	2.61	1.90	1.33	0.78	0.47	1.12	1.55	1.50	1.40	1.30	1.18	0.20
PC4	1.82	1.31	0.81	0.41	0.21	0.89	1.02	0.97	0.90	0.82	0.78	0.14
Double-cross hybrid DKB747 (n=416 plants) in the 2008/2009 crop year												
PC1	10.07	8.69	6.81	4.98	3.76	3.72	7.19	6.98	6.74	6.48	6.24	0.50
PC2	3.92	2.92	2.03	1.26	0.63	1.65	1.71	1.61	1.49	1.38	1.29	0.23
PC3	2.42	1.97	1.34	0.77	0.51	1.20	1.41	1.33	1.23	1.13	1.08	0.21
PC4	1.80	1.36	0.85	0.43	0.25	0.92	1.23	1.14	1.05	0.97	0.87	0.17
Single-cross hybrid 30F53 (n=1,777 plants) in the 2009/2010 crop year												
PC1	9.49	8.69	7.51	6.06	4.49	2.63	7.57	7.48	7.31	7.14	7.00	0.34
PC2	3.36	2.59	1.88	1.36	0.93	1.23	1.77	1.67	1.60	1.53	1.48	0.14
PC3	2.76	1.69	1.17	0.68	0.39	1.00	1.32	1.25	1.17	1.09	1.02	0.16
PC4	1.37	1.10	0.68	0.34	0.18	0.76	0.88	0.78	0.70	0.62	0.57	0.16
Three-way cross hybrid DKB566 (n=1,693 plants) in the 2009/2010 crop year												
PC1	9.30	8.37	6.93	5.35	4.25	3.03	7.03	6.89	6.70	6.52	6.41	0.38
PC2	4.03	2.85	2.03	1.40	1.01	1.45	1.78	1.72	1.67	1.61	1.56	0.12
PC3	2.29	1.86	1.27	0.73	0.39	1.12	1.42	1.35	1.26	1.16	1.09	0.19
PC4	1.58	1.26	0.78	0.41	0.23	0.85	0.94	0.87	0.79	0.72	0.66	0.15
Double-cross hybrid DKB747 (n=1,720 plants) in the 2009/2010 crop year												
PC1	9.20	8.29	6.78	5.13	3.64	3.17	7.03	6.75	6.55	6.35	6.18	0.41
PC2	4.17	2.94	2.08	1.39	1.05	1.55	1.87	1.80	1.70	1.60	1.54	0.20
PC3	2.38	1.92	1.33	0.79	0.56	1.13	1.50	1.39	1.29	1.19	1.13	0.20
PC4	1.71	1.30	0.83	0.44	0.28	0.86	1.05	0.98	0.89	0.81	0.75	0.17

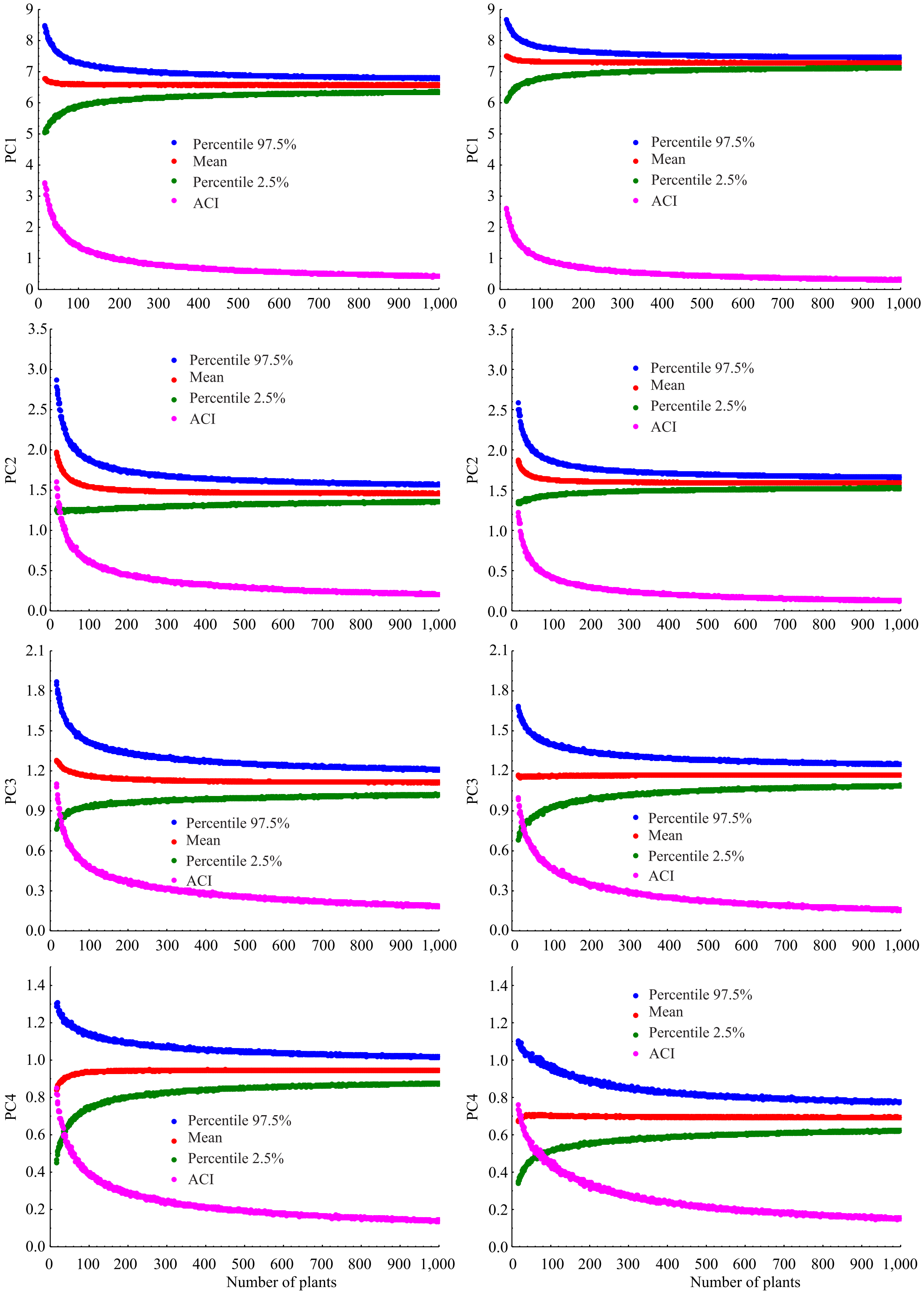

Figure 1. Percentile 97.5%, mean, percentile 2.5%, and amplitude of confidence interval of 95% (ACI) for 3,000 estimates of eigenvalues of the first four principal components (PC1, PC2, PC3, and PC4), based in resampling of the 361 plants of single-cross hybrid P32R21 in the 2008/2009 crop year (left column), and 1,777 plants of single-cross hybrid 30F53 in the 2009/2010 crop year (right column). On the X axis the number of plants ranges from 15 to 1,000.

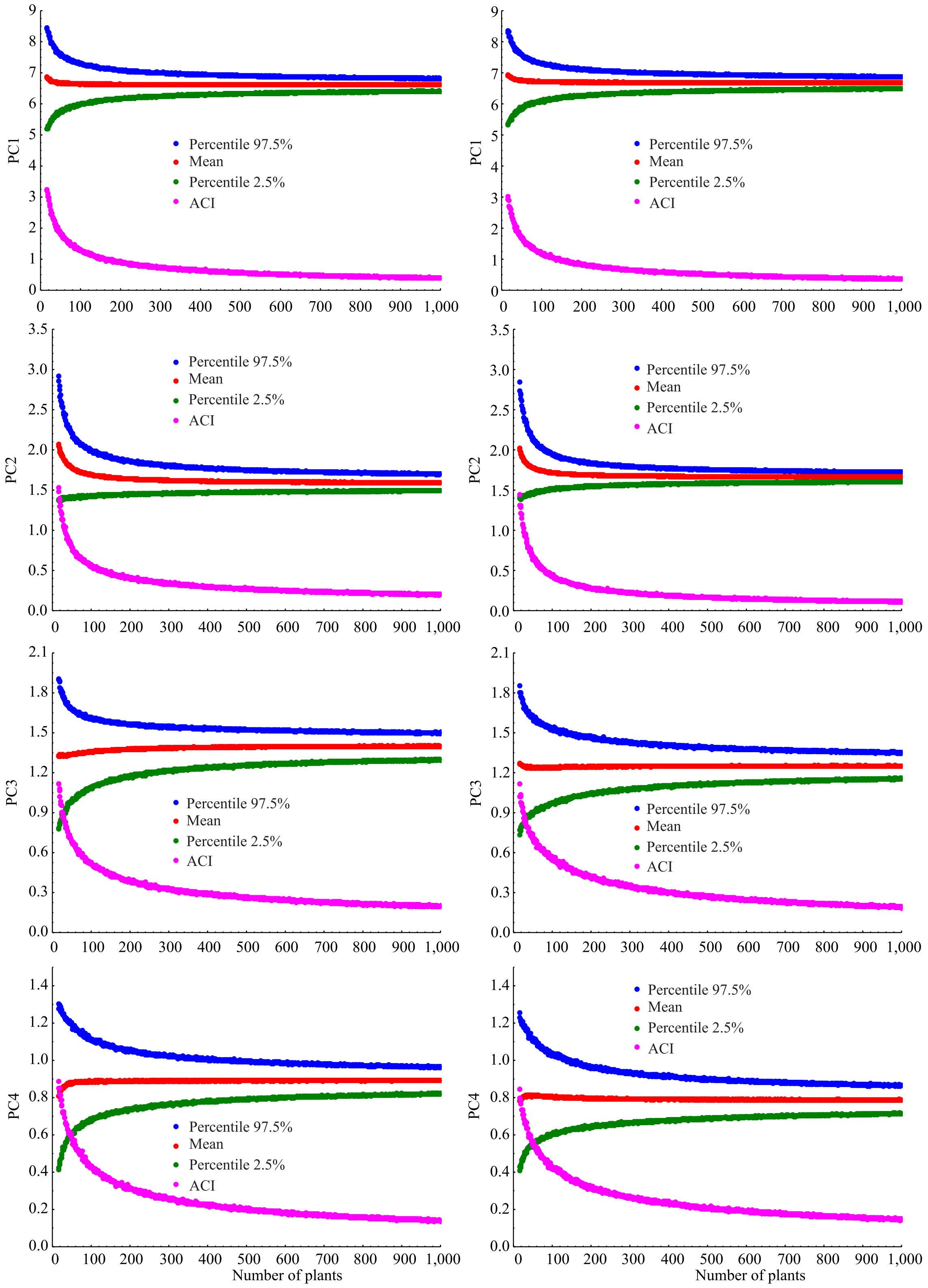

Figure 2. Percentile 97.5%, mean, percentile 2.5%, and amplitude of confidence interval of 95% (ACI) for 3,000 estimates of eigenvalues of the first four principal components (PC1, PC2, PC3, and PC4), based in resampling of the 373 plants of three-way cross hybrid DKB566 in the 2008/2009 crop year (left column) and 1,693 plants of three-way cross hybrid DKB566 in the 2009/2010 crop year (right column). On the X axis the number of plants ranges from 15 to 1,000.

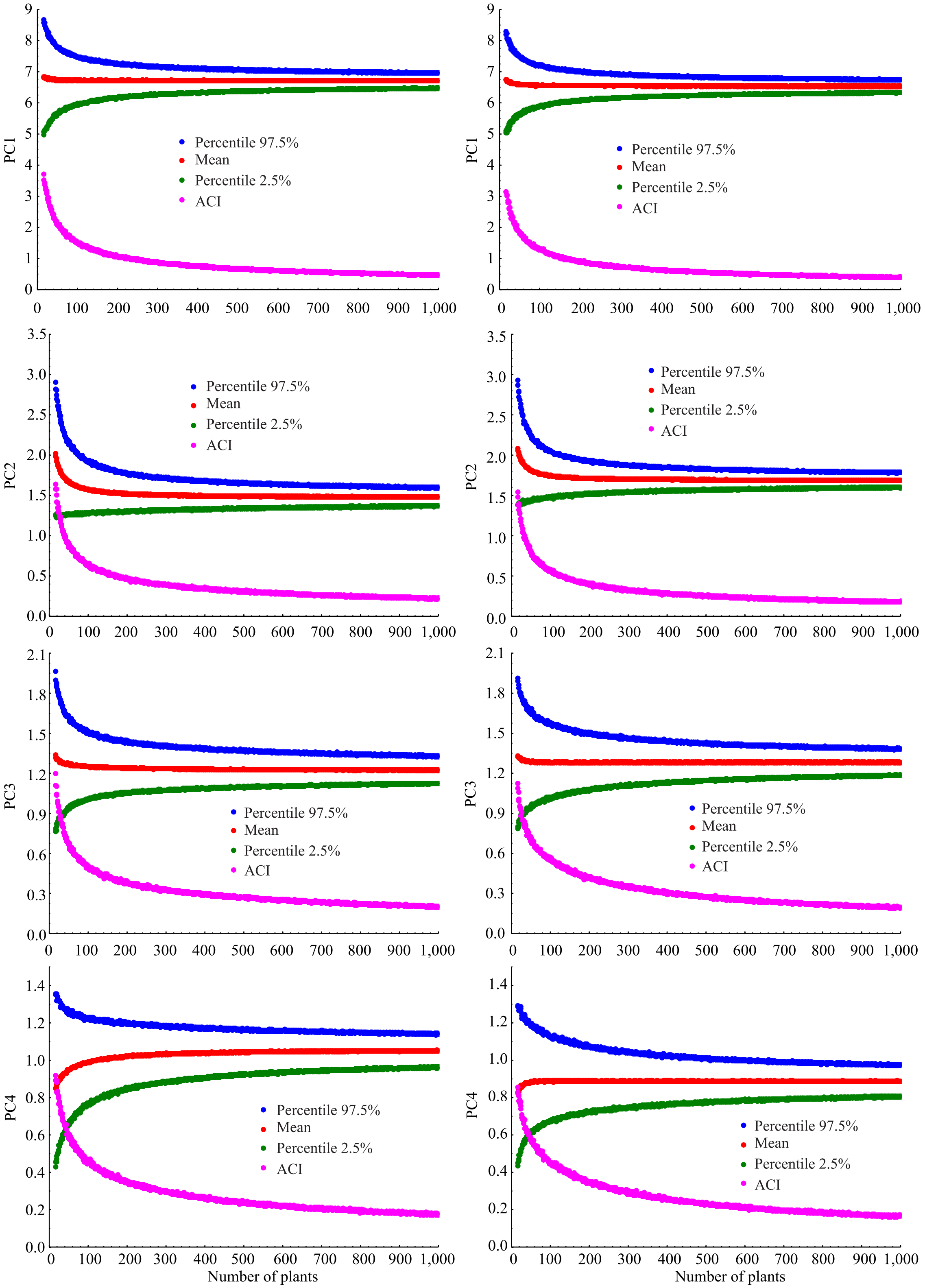

Figure 3. Percentile 97.5%, mean, percentile 2.5%, and amplitude of confidence interval of 95% (ACI) for 3,000 estimates of eigenvalues of the first four principal components (PC1, PC2, PC3, and PC4), based in resampling of the 416 plants of double-cross hybrid DKB747 in the 2008/2009 crop year (left column) and 1,720 plants of double-cross hybrid DKB747 in the 2009/2010 crop year (right column). On the X axis the number of plants ranges from 15 to 1,000.

multiple regression models of corn grain yield as a function of ear length and ear diameter. Farther, the information from the present research includes the sample size for principal components analysis, which was not explored in previous studies on corn crop or another agricultural crop.

Ramachandran & Aschheim (2005) identified that scenarios with principal components with less explanatory variance required a larger sample size (n = 100) than that in data with principal components with high explanatory variance (n = 20). These authors also identified that when the sample size increases, the errors become small and finally reach a constant value from a certain simulated sample size. Osborne & Costello (2004) observed an interaction between the number of observations and the subject to item ratio, and the best results were obtained under conditions of high number of observations and subject to item ratios. Kocovsky et al. (2009) observed that in small sample sizes, eigenvalues for the first principal components were unstable and inflated, and they recommended a minimum subject to item ratio from 3.5 to 8.0 to

Table 3. Parameter estimates of the model linear response with plateau (a, b), determination coefficient (r^2), required sample size (n_o, number of plants) to estimate the eigenvalues of the first four principal components (PC1, PC2, PC3, and PC4), and amplitude of confidence interval of 95% in sample size ACI(n_o) for corn hybrids (*Zea mays*) grown in two crop years.

PC	a	b	r^2	n_o	ACI(n_o)
Single-cross hybrid P32R21 (n=361 plants) in the 2008/2009 crop year					
PC1	2.41629	-0.00770	0.870	236	0.598
PC2	1.05305	-0.00324	0.848	238	0.282
PC3	0.73308	-0.00179	0.849	274	0.241
PC4	0.59689	-0.00150	0.876	278	0.181
Three-way cross hybrid DKB566 (n=373 plants) in the 2008/2009 crop year					
PC1	2.28036	-0.00751	0.867	230	0.552
PC2	1.03608	-0.00376	0.851	204	0.269
PC3	0.78876	-0.00198	0.872	272	0.250
PC4	0.64283	-0.00156	0.887	294	0.185
Double-cross hybrid DKB747 (n=416 plants) in the 2008/2009 crop year					
PC1	2.57669	-0.00792	0.872	243	0.652
PC2	1.10661	-0.00352	0.851	229	0.300
PC3	0.80294	-0.00226	0.851	241	0.258
PC4	0.65355	-0.00139	0.875	311	0.221
Single-cross hybrid 30F53 (n=1,777 plants) in the 2009/2010 crop year					
PC1	1.80228	-0.00590	0.862	231	0.439
PC2	0.81845	-0.00309	0.852	205	0.185
PC3	0.69493	-0.00158	0.882	308	0.208
PC4	0.57522	-0.00104	0.915	372	0.190
Three-way cross hybrid DKB566 (n=1,693 plants) in the 2009/2010 crop year					
PC1	2.09806	-0.00683	0.867	232	0.513
PC2	0.98483	-0.00472	0.860	172	0.172
PC3	0.76152	-0.00150	0.894	344	0.245
PC4	0.60129	-0.00127	0.890	324	0.190
Double-cross hybrid DKB747 (n=1,720 plants) in the 2009/2010 crop year					
PC1	2.25906	-0.00728	0.870	234	0.555
PC2	1.00986	-0.00337	0.855	226	0.248
PC3	0.75625	-0.00148	0.884	344	0.248
PC4	0.60836	-0.00111	0.890	360	0.208
Overall mean					
PC1	-[1]	-	0.868	234	0.552
PC2	-	-	0.853	212	0.243
PC3	-	-	0.872	297	0.242
PC4	-	-	0.889	323	0.196

[1] Overall mean not calculated.

increase the eigenvalues and eigenvectors stabilization. Although 267 plants from the present study seem to be a large number of observations, which could discourage the application of the PCA technique in corn by some researchers, it is important to highlight the indication of Osborne & Costello (2004) that researchers need to remember that PCA is a large-sample technique, not well-suited for the small sample sizes some researchers employ. Furthermore, these authors emphasize that in many cases the subject to item ratio should be greater than 20:1 and the sample size greater than 1,000.

Conclusion

The measurement of 267 plants is sufficient to estimate the eigenvalues of the principal components for corn (*Zea mays*) traits.

Acknowledgments

To Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), for research grant to the first author (process number 304652/2017-2); and to those who assisted in carrying out the experiment and in data collection.

References

ALI, F.; KANWAL, N.; AHSAN, M.; ALI, Q.; BIBI, I.; NIAZI, N.K. Multivariate analysis of grain yield and its attributing traits in different maize hybrids grown under heat and drought stress. **Scientifica**, v.2015, art.563869, 2015. DOI: https://doi.org/10.1155/2015/563869.

ALVARES, C.A.; STAPE, J.L.; SENTELHAS, P.C.; GONÇALVES, J.L. de M.; SPAROVEK, G. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v.22, p.711-728, 2013. DOI: https://doi.org/10.1127/0941-2948/2013/0507.

BELALIA, N.; LUPINI, A.; DJEMEL, A.; MORSLI, A.; MAUCERI, A.; LOTTI, C.; KHELIFI-SLAOUI, M.; KHELIFI, L.; SUNSERI, F. Analysis of genetic diversity and population structure in Saharan maize (*Zea mays* L.) populations using phenotypic traits and SSR markers. **Genetic Resources and Crop Evolution**, v.66, p.243-257, 2019. DOI: https://doi.org/10.1007/s10722-018-0709-3.

BJÖRKLUND, M. Be careful with your principal components. **Evolution**, v.73, p.2151-2158, 2019. DOI: https://doi.org/10.1111/evo.13835.

CARGNELUTTI FILHO, A.; TOEBE, M. Reference sample size for multiple regression in corn. **Pesquisa Agropecuária Brasileira**, v.55, e01400, 2020. DOI: https://doi.org/10.1590/s1678-3921.pab2020.v55.01400.

CARGNELUTTI FILHO, A.; TOEBE, M.; BURIN, C.; SILVEIRA, T.R. da; CASAROTTO, G. Tamanho de amostra para estimação do coeficiente de correlação linear de Pearson entre caracteres de milho. **Pesquisa Agropecuária Brasileira**, v.45, p.1363-1371, 2010. DOI: https://doi.org/10.1590/S0100-204X2010001200005.

DOCHTERMANN, N.A.; JENKINS, S.H. Multivariate methods and small sample sizes. **Ethology**, v.117, p.95-101, 2011. DOI: https://doi.org/10.1111/j.1439-0310.2010.01846.x.

FANCELLI, A.L.; DOURADO NETO, D. (Ed.). **Milho**: manejo e produtividade. Piracicaba: ESALQ/USP, 2009. 181p.

FERREIRA, D.F. **Estatística básica**. 2.ed. rev. Lavras: UFLA. 2009. 664p.

FERREIRA, D.F. **Estatística multivariada**. 3.ed. Lavras: UFLA, 2018. 624p.

GAÑAN-CARDENAS, E.; CORREA-MORALES, J.C. Comparison of correction factors and sample size required to test the equality of the smallest eigenvalues in principal component analysis. **Revista Colombiana de Estadística**, v.44, p.43-64, 2021.

GUIMARÃES, P. de S.; BERNINI, C.S.; PEDROSO, F.K.J.V.; PATERNIANI, M.E.A.G.Z. Characterizing corn hybrids (*Zea mays* L) for water shortage by principal components analysis. **Maydica**, v.59, p.72-79, 2014.

KOCOVSKY, P.M.; ADAMS, J.V.; BRONTE, C.R. The effect of sample size on the stability of principal components analysis of truss-based fish morphometrics. **Transactions of the American Fisheries Society**, v.138, p.487-496, 2009. DOI: https://doi.org/10.1577/T08-091.1.

LATTIN, J.; CARROLL, J.D.; GREEN, P.E. **Análise de dados multivariados**. São Paulo: Cengage Learning, 2011. 475p.

OLIVOTO, T.; LÚCIO, A.D.; SOUZA, V.Q. de; NARDINO, M.; DIEL, M.I.; SARI, B.G.; KRYSCZUN, D.K.; MEIRA, D.; MEIER, C. Confidence interval width for Pearson's correlation coefficient: a Gaussian-independent estimator based on sample size and strength of association. **Agronomy Journal**, v.110, p.503-510, 2018. DOI: https://doi.org/10.2134/agronj2017.09.0566.

OLIVOTO, T.; NARDINO, M.; CARVALHO, I.R.; FOLLMANN, D.N.; FERRARI, M.; PELEGRIN, A.J. de; SZARESKI, V.J.; OLIVEIRA, A.C. de; CARON, B.O.; SOUZA, V.Q. de. 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**, v.12, p.93-103, 2017. DOI: https://doi.org/10.5897/AJAR2016.11799.

OSBORNE, J.W.; COSTELLO, A.B. Sample size and subject to item ratio in principal components analysis. **Practical Assessment, Research and Evaluation**, v.9, art.11, 2004. DOI: https://doi.org/10.7275/ktzq-jq66.

PARANAÍBA, P.F.; FERREIRA, D.F.; MORAIS, A.R. de. Tamanho ótimo de parcelas experimentais: proposição de métodos de estimação. **Revista Brasileira de Biometria**, v.27, p.255-268, 2009.

R CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2021. Available at: <http://www.R-project.org>. Accessed on: Mar. 15 2021.

RAMACHANDRAN, J.; ASCHHEIM, M.A. Sample size and error in the determination of mode shapes by principal components analysis. **Engineering Structures**, v.27, p.1951-1967, 2005. DOI: https://doi.org/10.1016/j.engstruct.2005.06.020.

SANTOS, H.G. dos; JACOMINE, P.K.T.; ANJOS, L.H.C. dos; OLIVEIRA, V.Á. de; LUMBRERAS, J.F.; COELHO, M.R.; ALMEIDA, J.A. de; ARAÚJO FILHO, J.C. de; OLIVEIRA, J.B. de; CUNHA, T.J.F. **Sistema brasileiro de classificação de solos**. 5.ed. rev. e ampl. Brasília: Embrapa, 2018. 356p.

SHAUKAT, S.S.; RAO, T.A.; KHAN, M.A. Impact of sample size on principal component analysis ordination of an environmental data set: effects on eigenstructure. **Ekológia Bratislava**, v.35, p.173-190, 2016. DOI: https://doi.org/10.1515/eko-2016-0014.

SOIL SURVEY STAFF. **Soil taxonomy**: a basic system of soil classification for making and interpreting soil surveys. 2nd ed. Washington: USDA, NRCS, 1999. 886p. (Agriculture Handbook, 436).

STAUFFER, D.F.; GARTON, E.O.; STEINHORST, R.K. A comparison of principal components from real and random data. **Ecology**, v.66, p.1693-1698, 1985. DOI: https://doi.org/10.2307/2937364.

TOEBE, M.; CARGNELUTTI FILHO, A.; BURIN, C.; CASAROTTO, G.; HAESBAERT, F.M. Tamanho de amostra para estimação da média e do coeficiente de variação em milho. **Pesquisa Agropecuária Brasileira**, v.49, p.860-871, 2014. DOI: https://doi.org/10.1590/S0100204X2014001100005.

TOEBE, M.; CARGNELUTTI FILHO, A.; LOPES, S.J.; BURIN, C.; SILVEIRA, T.R. da; CASAROTTO, G. Sample size in the estimation of correlation coefficients for corn hybrids in crops and accuracy levels. **Bragantia**, v.74, p.16-24, 2015. DOI: https://doi.org/10.1590/1678-4499.0324.

TOEBE, M.; CARGNELUTTI FILHO, A.; STORK, L.; LÚCIO, A.D. Sample size for estimation of direct effects in path analysis of corn. **Genetics and Molecular Research**, v.16, gmr16029523, 2017. DOI: https://doi.org/10.4238/gmr16029523.

WU, G.; MILLER, N.D.; DE LEON, N.; KAEPPLER, S.M.; SPALDING, E.P. Predicting *Zea mays* flowering time, yield, and kernel dimensions by analyzing aerial images. **Frontiers in Plant Science**, v.10, e1251, 2019. DOI: https://doi.org/10.3389/fpls.2019.01251.