CROP PRODUCTION AND MANAGEMENT
https://doi.org/10.1590/1678-4499.20210335 Article

Sample size for canonical correlation analysis in corn

Alberto Cargnelutti Filho[1,*], Marcos Toebe[2]

1. Universidade Federal de Santa Maria – Departamento de Fitotecnia – Santa Maria (RS), Brazil.
2. Universidade Federal de Santa Maria – Departamento de Fitotecnia – Frederico Westphalen (RS), Brazil.

Received: Dec. 3, 2021 | **Accepted:** June 7, 2022

Section Editor: Gabriel Constantino Blain

***Corresponding author:** alberto.cargnelutti.filho@gmail.com

How to cite: Cargnelutti Filho, A. and Toebe, M. (2022). Sample size for canonical correlation analysis in corn. Bragantia, 81, e3722.https://doi.org/10.1590/1678-4499.20210335

ABSTRACT: The canonical correlation analysis has been successfully used in many areas aiming to extract important information from a pair of data sets. Thus, the objective of this work was to determine the sample size (number of plants) required to estimate the canonical correlations in corn characteristics. Six characteristics were measured in 361, 373, and 416 plants, respectively, of the single, three-way and double cross hybrids of the 2008/2009 crop year and in 1,777, 1,693, and 1,720 plants, respectively, of the single, three-way, and double cross hybrids (2009/2010 crop) (six cases). The canonical correlation analyses were carried out between characteristics group of the plant architecture (plant height at harvest and ear insertion height) *versus* grain production (hundred grains mass and grains mass per plant) (scenario 1), and dimensions of ear (ear length and ear diameter) *versus* grain production (hundred grains mass and grains mass per plant) (scenario 2). The sample size (number of plants) for the estimation of canonical correlations was determined by resampling with replacement and application of the model linear response with plateau. Measuring 270 plants is sufficient to estimate the canonical correlation between groups with two characteristics in each group for corn. This sample size can be used as reference for reliable canonical correlation analysis.

Key words: *Zea mays*, multivariate analysis, resampling, model linear response with plateau.

INTRODUCTION

Corn is the cereal with the highest volume of world production, being one of the main inputs used in the animal protein production. The crop has been the focus of intense research aimed at improving productivity by area and nutritional components in grains and plants. Therefore, multiple variables and treatments are evaluated, and univariate and multivariate analyses are applied in order to improve decision making in relation to the best treatments and to verify the existing interrelationships between variables.

One technique of data analysis used in order to characterize the relations between two sets of variates is the canonical correlation analysis (CCA) developed by Hotelling (1936). According to Leach and Henson (2014), Uurtio et al. (2017) and Wu and Li (2021), the CCA has been successfully used in many areas aiming to extract important information from a pair of data sets, maximizing linear combinations between two sets of variables (characteristics). CCA has been used to identify the relationship between primary and secondary components of corn yield (Ceccon et al. 2016) and to assess the relationship between agronomic, protein-nutritional, and energetic-nutritional characteristics in corn genotypes (Alves et al. 2016). Also in corn genotypes, Alves et al. (2017) evaluated the effects of multicollinearity under two methods of CCA (with and without elimination of variables), and Crevelari et al. (2019) used CCA to study the linear dependence between groups of morphoagronomic and bromatological characteristics in silage corn hybrids.

CCA, like any other data analysis technique, is influenced by the original data set, which in this case is used to initially calculate the correlation matrix and, from this, the canonical correlations, as well as the significance of the correlations between groups of variables. As the measurement of all elements of the population is usually not feasible or possible, experiments are carried out, and samples are obtained to calculate the correlations. It is known that in the sampling process

there will always be an associated margin of error. Thus, the smaller the sample size used, the greater the inaccuracy in the estimates of correlations and complementary analyses, such as the canonical correlation.

Some studies in CCA have already pointed out the problem of small sample size (SSS), especially when the dimensionality of the data is greater than the number of observations. In this sense, authors such as Barcikowski and Stevens (1975), Thompson (1990), Sun et al. (2010), Leach and Henson (2014), Song et al. (2016), Uurtio et al. (2017), Krzyśko et al. (2018), Helmer et al. (2021) and Wu and Li (2021) pointed out alternatives, established number of subjects per variable and/or proposals for complementary or alternative methods to solve SSS problem for carrying out the analysis via CCA. Some of the aforementioned authors also indicated that the CCA performed under SSS are not only imprecise, but often overfitting the strength of the association between groups of variables.

Therefore, the question to be answered in this work was: what is the minimum number of plants that must be evaluated for the accurate estimate of CCA in corn. So, the objective of this work was to determine the sample size (number of plants) required to estimate the canonical correlations in corn characteristics.

MATERIALS AND METHODS

Data from two experiments with corn (*Zea mays* L.) were used. The experiments were carried out in the 2008/2009 (first experiment) and 2009/2010 (second experiment) crop year, in an experimental area located at 29º42'S, 53º49'W, 95 m altitude, Santa Maria, state of Rio Grande do Sul, Brazil.

In the first experiment, 361 plants of the single cross hybrid P32R21 (case 1); 373 plants of the three-way cross hybrid DKB566 (case 2); and 416 plants of the double cross hybrid DKB747 (case 3) were evaluated. In the second experiment, 1,777 of the single cross hybrid 30F53 (case 4); 1,693 plants of the three-way cross hybrid DKB566 (case 5); and 1,720 plants of the double cross hybrid DKB747 (case 6) were evaluated. In these 6,340 plants, the following characteristics were measured: plant height at harvest (PH, in cm), ear insertion height (EIH, in cm), ear length (EL, in cm), ear diameter (ED, in mm), hundred grains mass (HGM, in g), and grains mass per plant (GM, in g per plant).

The CCA was performed for each hybrid in each experiment (six cases), from the Pearson's linear correlation matrix, by the CCA function of MVar.pt package of software R (R Core Team 2021). The CCA was carried out between characteristics group of the plant architecture (plant height at harvest and ear insertion height) *versus* grain production (hundred grains mass and grains mass per plant) (scenario 1), and dimensions of ear (ear length and ear diameter) *versus* grain production (hundred grains mass and grains mass per plant) (scenario 2). The diagnosis of multicollinearity was performed by condition number (CN) between characteristics of each group.

The sample size (n_o, number of plants) required to estimate the canonical correlation was determined through resampling with replacement. For resampling, 991 sample sizes were planned, with an initial sample size of ten plants (in this study, considered as the minimum size required for CCA). The other sizes were obtained in increments of one unit, until reaching 1,000 plants. Thus, sample sizes of 10 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 first (CC1) and second (CC2) canonical correlation were obtained. Thus, for each sample size, 3,000 estimates of the CC1 and CC2 were obtained, and the percentile 97.5% ($P_{97.5\%}$), mean, and percentile 2.5% ($P_{2.5\%}$) were determined. The amplitude of confidence interval of 95% was calculated by Eq. 1:

$$ACI = P_{97.5\%} - P_{2.5\%} \quad (1)$$

It should be interpreted that the smaller the ACI, the more accurate are the estimates of the first and second canonical correlation.

For six cases and two scenarios, the sample size (n_o, number of plants) required to estimate the CC1 and CC2 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 parameters a, b and p 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 slope (b) not null. The second straight [$ACI_{(n)} = p + \varepsilon$] started from n_o and had a zero slope, that is, it was 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 $n_o = (p\text{-}a)/b$ and the amplitude of the confidence interval in the optimal sample size by $ACIn_o = a + bn_o$.

The percentile 97.5% ($P_{97.5\%}$), mean, percentile 2.5% ($P_{2.5\%}$), and amplitude of confidence interval of 95% (ACI) for n = 10 and n = 1,000 plants were presented in a table, and the other ones 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

The condition number of correlation matrix fluctuated among 3.77 and 7.44 for characteristics of the plant architecture (PH and EIH), among 3.58, and 7.42 for characteristics of the dimensions of ear (EL and ED); and among 2.55 and 3.75 for characteristics of the grain production (HGM and GM). These results indicate that in the original data there was no problem of collinearity within the evaluated groups.

The first canonical correlation (CC1) between characteristics of the plant architecture (PH and EIH) and grain production (HGM and GM) (scenario 1) was significant and of intermediate magnitude in the six cases. It oscillated between 0.3127 and 0.6158, with average of 0.4194, showing that the groups are dependent (Table 1). It is interpreted, through canonical coefficients, that the tallest plants are associated with the highest mass of grains per plant and vice versa. Despite the statistical significance of the second canonical correlation (CC2) in four of the six cases, the practical significance is negligible due to the low magnitude (0.0077 ≤ canonical correlation ≤ 0.1228).

The CC1 between the characteristics of the dimensions of ear (EL and ED) and grain production (HGM and GM) (scenario 2) was significant and of high magnitude in the six cases. It ranged between 0.8875 and 0.9487, with average of 0.9163, showing that the groups are dependent (Table 1). It is interpreted, through canonical coefficients, that plants with longer ears and larger diameter are associated with greater mass of grains per plant. Despite the statistical significance of CC2 in four of the six cases, the practical significance is negligible due to the low magnitude (0.0496 ≤ canonical correlation ≤ 0.2597).

In the CC1 between characteristics of the plant architecture (PH and EIH) and grain production (HGM and GM) (scenario 1), from the 3,000 resamples of 10 plants (the smallest sample size used in this study), the 95% confidence interval (ACI) was 0.6436, 0.6331, 0.6135, 0.5424, 0.6014, and 0.6260, and the mean of the 3,000 resamples was 0.6742, 0.6596, 0.6890, 0.7577, 0.6658, and 0.6371, respectively, for the cases 1, 2, 3, 4, 5, and 6 (Table 2 and Fig. 1). At the other end, from the 3,000 resamples from 1,000 plants (largest sample size used), the ACI was 0.1252, 0.1244, 0.1083, 0.0863, 0.1143, and 0.1171, and the mean of the 3,000 resamples was 0.4006, 0.3498, 0.4855, 0.6178, 0.3632, and 0.3157, respectively, for the cases 1, 2, 3, 4, 5, and 6. Visually, it can be seen that, with the increase in the number of plants, the mean of the 3,000 estimates of the CC1, in the six cases, stabilizes and approaches the obtained with the 361 plants of the single cross hybrid P32R21 (case 1 – CC1 = 0.3986), 373 plants of the three-way cross hybrid DKB566 (case 2 – CC1 = 0.3464), 416 plants of the double cross hybrid DKB747 (case 3 – CC1 = 0.4834), 1,777 plants of the simple cross hybrid 30F53 (case 4 – CC1 = 0.6158), 1,693 plants of the three-way hybrid DKB566 (case 5 – CC1 = 0.3596), and 1,720 plants of the double cross hybrid DKB747 (case 6 – CC1 = 0.3127).

Table 1. Canonical correlations and coefficients of canonical pairs between characteristics of the plant architecture (PH: plant height at harvest, and EIH: ear insertion height), dimensions of ear (EL: ear length, and ED: ear diameter), and grain production (HGM: hundred grains mass, and GM: grains mass per plant) of corn hybrids (*Zea mays* L.), grown in two crop years.

	1st canonical pair	2nd canonical pair		1st canonical pair	2nd canonical pair
Single cross hybrid P32R21 (n = 361 plants) in the 2008/2009 crop year					
Correlation	0.3986*	0.0330ns	Correlation	0.9153*	0.2597*
Characteristics	Canonical coefficients		Characteristics	Canonical coefficients	
PH	-0.9044	0.8841	EL	0.5100	-1.0979
EIH	-0.1453	-1.2564	ED	0.6195	1.0401
HGM	-0.3298	-1.1398	HGM	-0.0167	1.1864
GM	-0.7831	0.8914	GM	1.0089	-0.6245
Three-way cross hybrid DKB566 (n = 373 plants) in the 2008/2009 crop year					
Correlation	0.3464*	0.1228*	Correlation	0.9158*	0.0496ns
Characteristics	Canonical coefficients		Characteristics	Canonical coefficients	
PH	-0.7144	1.3099	EL	0.5485	-1.2100
EIH	-0.3477	-1.4510	ED	0.5497	1.2095
HGM	-0.3731	1.0666	HGM	0.0810	1.1270
GM	-0.7702	-0.8268	GM	0.9597	-0.5964
Double cross hybrid DKB747 (n = 416 plants) in the 2008/2009 crop year					
Correlation	0.4834*	0.0077ns	Correlation	0.9198*	0.0624ns
Characteristics	Canonical coefficients		Characteristics	Canonical coefficients	
PH	-0.8600	1.1737	EL	-0.5945	1.1171
EIH	-0.1819	-1.4437	ED	-0.5185	-1.1543
HGM	-0.2449	-1.2014	HGM	-0.0077	-1.2261
GM	-0.8382	0.8949	GM	-0.9955	0.7157
Single cross hybrid 30F53 (n = 1,777 plants) in the 2009/2010 crop year					
Correlation	0.6158*	0.0853*	Correlation	0.9487*	0.1193*
Characteristics	Canonical coefficients		Characteristics	Canonical coefficients	
PH	-1.2004	-0.2622	EL	-0.6398	1.4073
EIH	0.9109	-0.8246	ED	-0.4225	-1.4870
HGM	-0.0547	1.1413	HGM	0.0801	1.1398
GM	-0.9724	-0.6000	GM	-1.0363	-0.4813
Three-way cross hybrid DKB566 (n = 1,693 plants) in the 2009/2010 crop year					
Correlation	0.3596*	0.1025*	Correlation	0.9111*	0.0840*
Characteristics	Canonical coefficients		Characteristics	Canonical coefficients	
PH	-1.3258	-0.2382	EL	0.4271	-1.1959
EIH	1.0650	-0.8247	ED	0.6785	1.0733
HGM	-0.5523	1.0469	HGM	-0.0243	-1.1834
GM	-0.5889	-1.0268	GM	1.0128	0.6127
Double cross hybrid DKB747 (n = 1,720 plants) in the 2009/2010 crop year					
Correlation	0.3127*	0.0229ns	Correlation	0.8875*	0.1838*
Characteristics	Canonical coefficients		Characteristics	Canonical coefficients	
PH	-1.3433	0.7676	EL	0.4895	-1.1796
EIH	0.5288	-1.4539	ED	0.6192	1.1169
HGM	-0.4983	0.9939	HGM	0.0841	-1.1086
GM	-0.6761	-0.8826	GM	0.9604	0.5601

*Significant by the χ^2 test at 5% probability of error; ns: not significant.

Table 2. Percentile 97.5% ($P_{97.5\%}$), mean, percentile 2.5% ($P_{2.5\%}$), and amplitude of confidence interval of 95% (ACI = $P_{97.5\%}$ - $P_{2.5\%}$) for 3,000 estimates of the first (CC1) and second (CC2) canonical correlation between characteristics of the plant architecture (plant height at harvest, and ear insertion height), dimensions of ear (ear length, and ear diameter), and grain production (hundred grains mass, and grains mass per plant). Estimates obtained from 3,000 resamples with replacement for n = 10 and 1,000 plants of corn hybrids (*Zea mays* L.), grown in two crop years.

Characteristics	Correlation	$P_{97.5\%}$	Mean	$P_{2.5\%}$	ACI	$P_{97.5\%}$	Mean	$P_{2.5\%}$	ACI
		---------- n = 10 plants ----------				---------- n = 1,000 plants ----------			
Single cross hybrid P32R21 in the 2008/2009 crop year									
Architecture *vs.* grains	CC1	0.9402	0.6742	0.2966	0.6436	0.4630	0.4006	0.3378	0.1252
Architecture *vs.* grains	CC2	0.5944	0.2280	0.0087	0.5857	0.0886	0.0365	0.0017	0.0869
Ear *vs.* grains	CC1	0.9914	0.9429	0.8227	0.1687	0.9251	0.9156	0.9051	0.0201
Ear *vs.* grains	CC2	0.7921	0.3664	0.0196	0.7725	0.3258	0.2597	0.1914	0.1344
Three-way cross hybrid DKB566 in the 2008/2009 crop year									
Architecture *vs.* grains	CC1	0.9237	0.6596	0.2907	0.6331	0.4136	0.3498	0.2892	0.1244
Architecture *vs.* grains	CC2	0.6113	0.2286	0.0091	0.6023	0.1823	0.1222	0.0641	0.1182
Ear *vs.* grains	CC1	0.9938	0.9431	0.7844	0.2094	0.9275	0.9161	0.9038	0.0237
Ear *vs.* grains	CC2	0.7567	0.3181	0.0136	0.7430	0.1114	0.0516	0.0034	0.1080
Double cross hybrid DKB747 in the 2008/2009 crop year									
Architecture *vs.* grains	CC1	0.9383	0.6890	0.3247	0.6135	0.5399	0.4855	0.4316	0.1083
Architecture *vs.* grains	CC2	0.6262	0.2442	0.0127	0.6135	0.0776	0.0272	0.0009	0.0767
Ear *vs.* grains	CC1	0.9917	0.9389	0.7812	0.2105	0.9291	0.9199	0.9099	0.0192
Ear *vs.* grains	CC2	0.7475	0.2940	0.0125	0.7350	0.1267	0.0638	0.0068	0.1199
Single cross hybrid 30F53 in the 2009/2010 crop year									
Architecture *vs.* grains	CC1	0.9563	0.7577	0.4138	0.5424	0.6593	0.6178	0.5730	0.0863
Architecture *vs.* grains	CC2	0.6592	0.2558	0.0095	0.6497	0.1436	0.0852	0.0253	0.1183
Ear *vs.* grains	CC1	0.9951	0.9655	0.8715	0.1236	0.9567	0.9490	0.9396	0.0171
Ear *vs.* grains	CC2	0.7795	0.3179	0.0136	0.7658	0.1900	0.1192	0.0468	0.1432
Three-way cross hybrid DKB566 in the 2009/2010 crop year									
Architecture *vs.* grains	CC1	0.9193	0.6658	0.3178	0.6014	0.4194	0.3632	0.3051	0.1143
Architecture *vs.* grains	CC2	0.6287	0.2358	0.0083	0.6204	0.1700	0.1013	0.0276	0.1425
Ear *vs.* grains	CC1	0.9908	0.9424	0.8242	0.1666	0.9212	0.9115	0.9003	0.0209
Ear *vs.* grains	CC2	0.7415	0.3102	0.0141	0.7274	0.1539	0.0842	0.0160	0.1379
Double cross hybrid DKB747 in the 2009/2010 crop year									
Architecture *vs.* grains	CC1	0.9131	0.6371	0.2871	0.6260	0.3757	0.3157	0.2586	0.1171
Architecture *vs.* grains	CC2	0.5667	0.2089	0.0062	0.5605	0.0823	0.0309	0.0016	0.0807
Ear *vs.* grains	CC1	0.9892	0.9251	0.7428	0.2465	0.9023	0.8878	0.8714	0.0309
Ear *vs.* grains	CC2	0.7669	0.3299	0.0131	0.7538	0.2480	0.1836	0.1162	0.1318

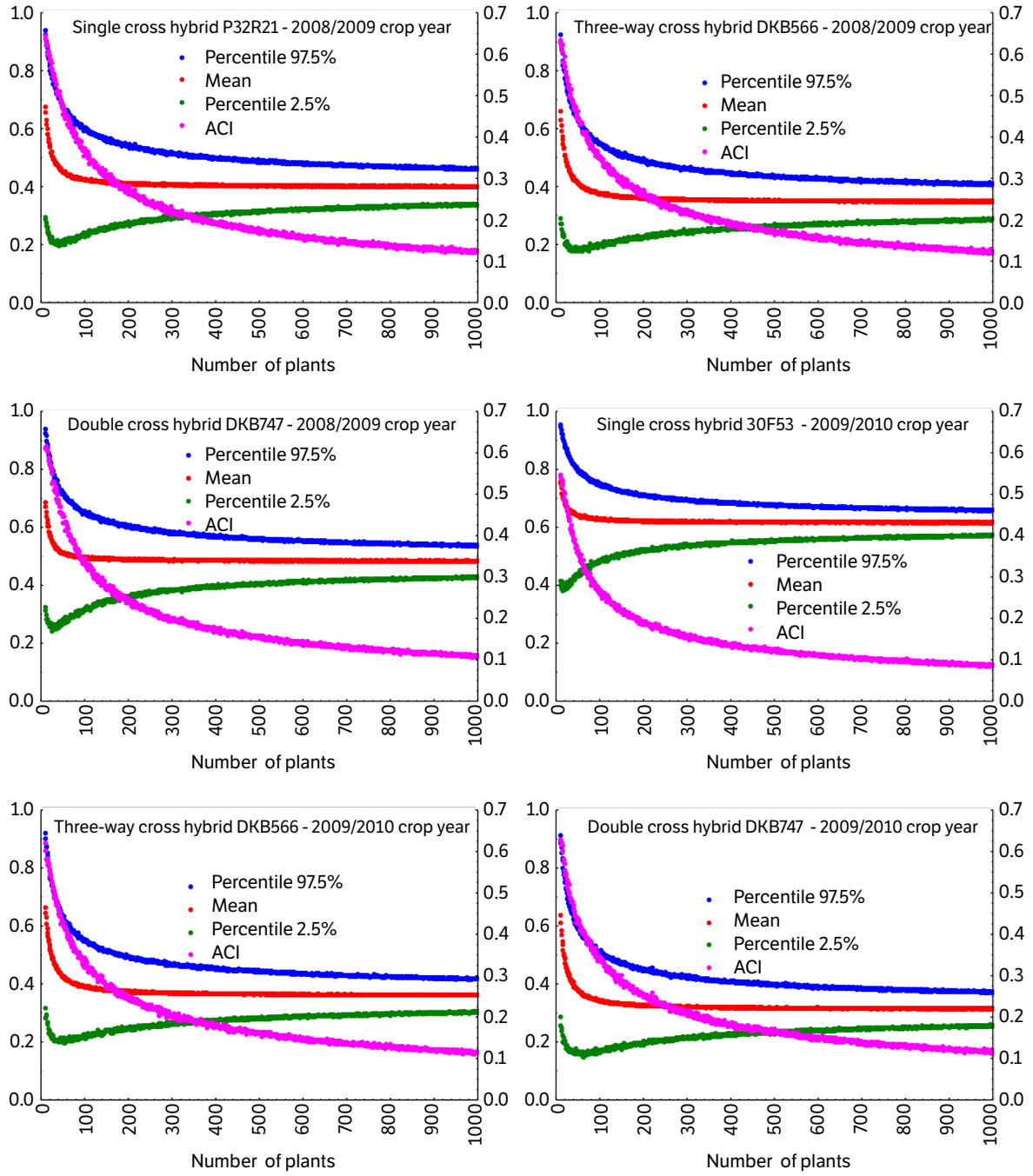

Figure 1. Percentile 97.5%, mean and percentile 2.5% (on the left Y-axis) and amplitude of confidence interval of 95% (ACI) (on the right Y-axis) for 3,000 estimates of first canonical correlation between characteristics of the plant architecture (plant height at harvest, and ear insertion height), and grain production (hundred grains mass, and grains mass per plant) of corn hybrids (*Zea mays* L.), grown in two crop years. On the X axis the number of plants ranges from 10 to 1,000.

A similar pattern was observed for CC2 between characteristics of the plant architecture (PH and EIH) and grain production (HGM and GM) (scenario 1) (Table 2 and Fig. 2) and for CC1 and CC2 between characteristics of the dimensions of ear (EL and ED) and grain production (HGM and GM) (scenario 2) (Table 2 and Figs. 3 and 4). Therefore, the greatest amplitudes of the confidence interval of the CC1 and CC2, in the six cases and two scenarios (architecture *vs.* grains and ear *vs.* grains), obtained from 10 plants compared to those obtained with 1,000 plants, show that with 10 plants the CC1 and CC2 estimates are less accurate, which may result in inaccurate and biased CCA, when the sample is insufficient.

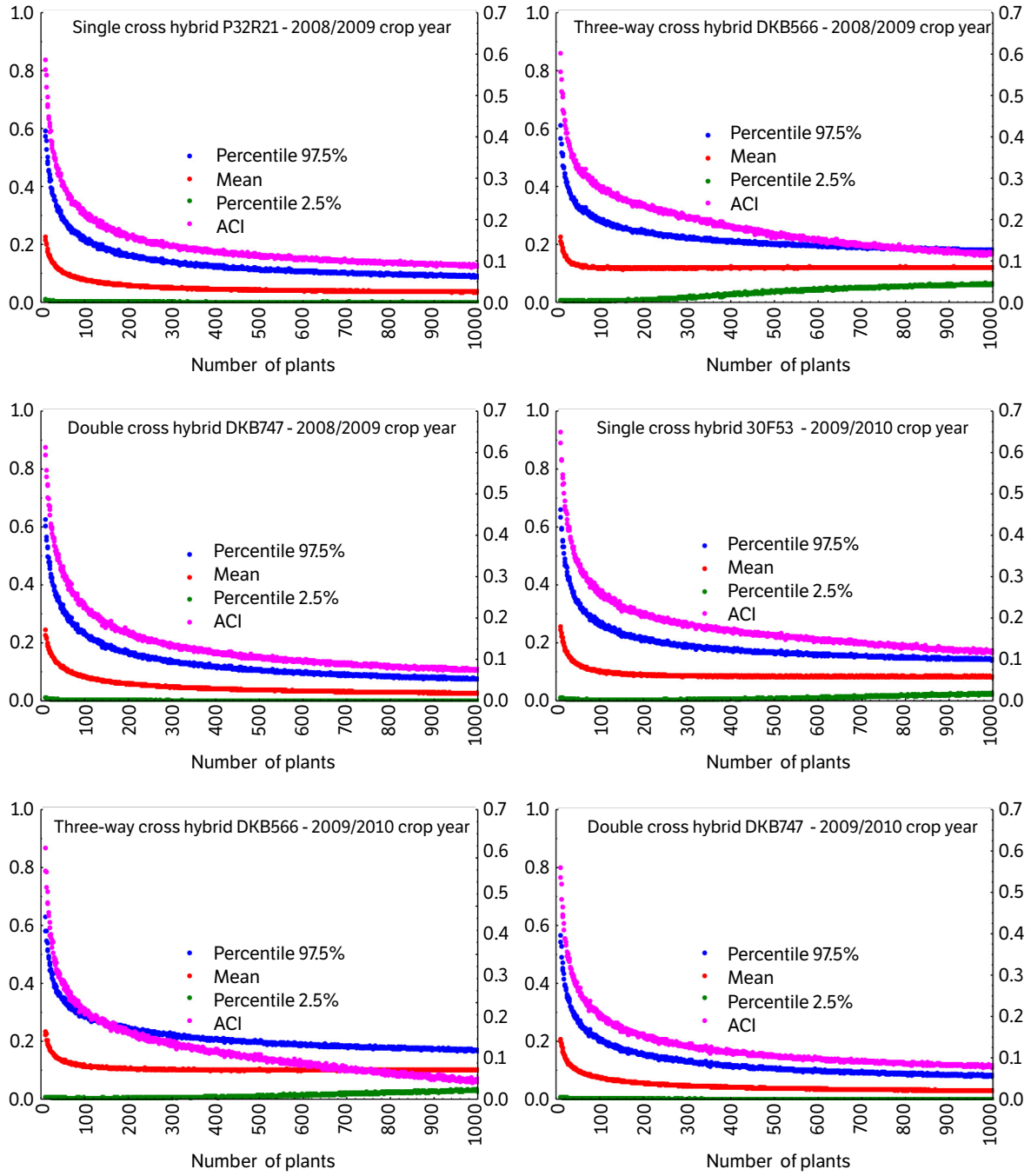

Figure 2. Percentile 97.5%, mean and percentile 2.5% (on the left Y-axis) and amplitude of confidence interval of 95% (ACI) (on the right Y-axis) for 3,000 estimates of second canonical correlation between characteristics of the plant architecture (plant height at harvest, and ear insertion height), and grain production (hundred grains mass, and grains mass per plant) of corn hybrids (*Zea mays* L.), grown in two crop years. On the X axis the number of plants ranges from 10 to 1,000.

The amplitudes of confidence interval of 95% (ACI), of the estimates of CC1 and CC2, in the six cases and two scenarios, gradually decreased with the increase in the number of plants (Figs. 1, 2, 3, and 4). Visually, it can be seen in Figs. 1, 2, 3, and 4 that there was a sharp decrease in the ACI up to approximately between 200 and 400 plants for the CC1 and CC2 for two scenarios, being possible to suggest that these sizes would be sufficient. Afterwards, the decreases are smaller, which indicates that the work to measure more plants would result in insignificant benefits in the precision of the estimates of the CC1 and CC2.

Based on model linear response with plateau, in the mean of the six cases, the samples size (n_o, number of plants) necessary to estimate the CC1 and CC2 for characteristics of the plant architecture (PH and EIH) and grain production (HGM and GM) (scenario 1) and CC1 and CC2 for characteristics of the dimensions of ear (EL and ED) and grain production (HGM and GM) (scenario 2) were, respectively, 311, 291, 212, and 263 plants (Table 3). Although estimates of the CC1 and CC2 from as many plants as possible should be aimed at guaranteeing reliable CCAs, it seems reasonable to estimate the canonical correlation based on 270 plants, which corresponds to the general mean of the 24 sample sizes (six cases × two scenarios × two canonical correlation). From this number of plants, the gains in precision (decrease in ACI) are insignificant (Figs. 1, 2, 3, and 4).

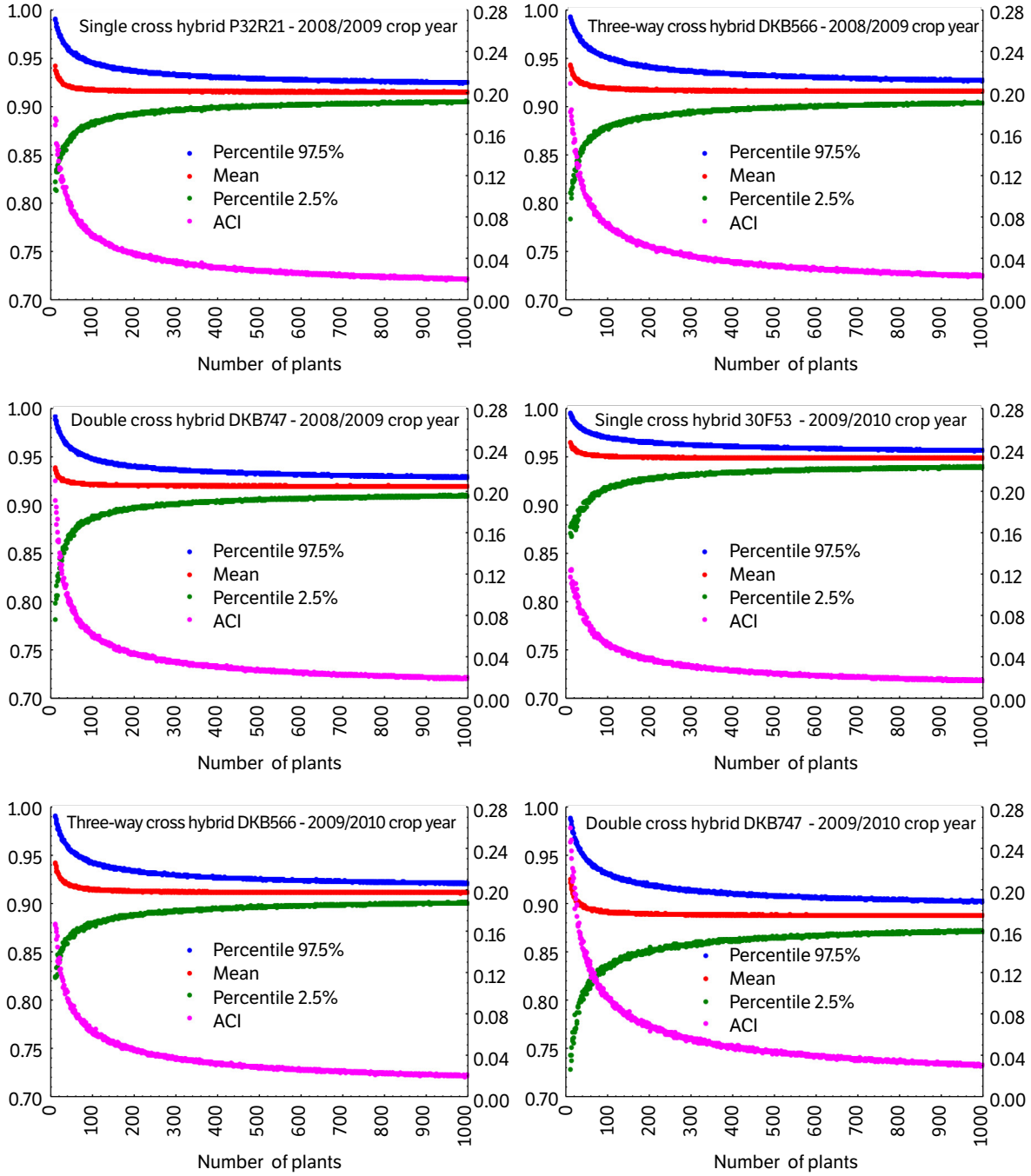

Figure 3. Percentile 97.5%, mean and percentile 2.5% (on the left Y-axis) and amplitude of confidence interval of 95% (ACI) (on the right Y-axis) for 3,000 estimates of first canonical correlation between characteristics of the dimensions of ear (ear length, and ear diameter), and grain production (hundred grains mass, and grains mass per plant) of corn hybrids (*Zea mays* L.), grown in two crop years. On the X axis the number of plants ranges from 10 to 1,000.

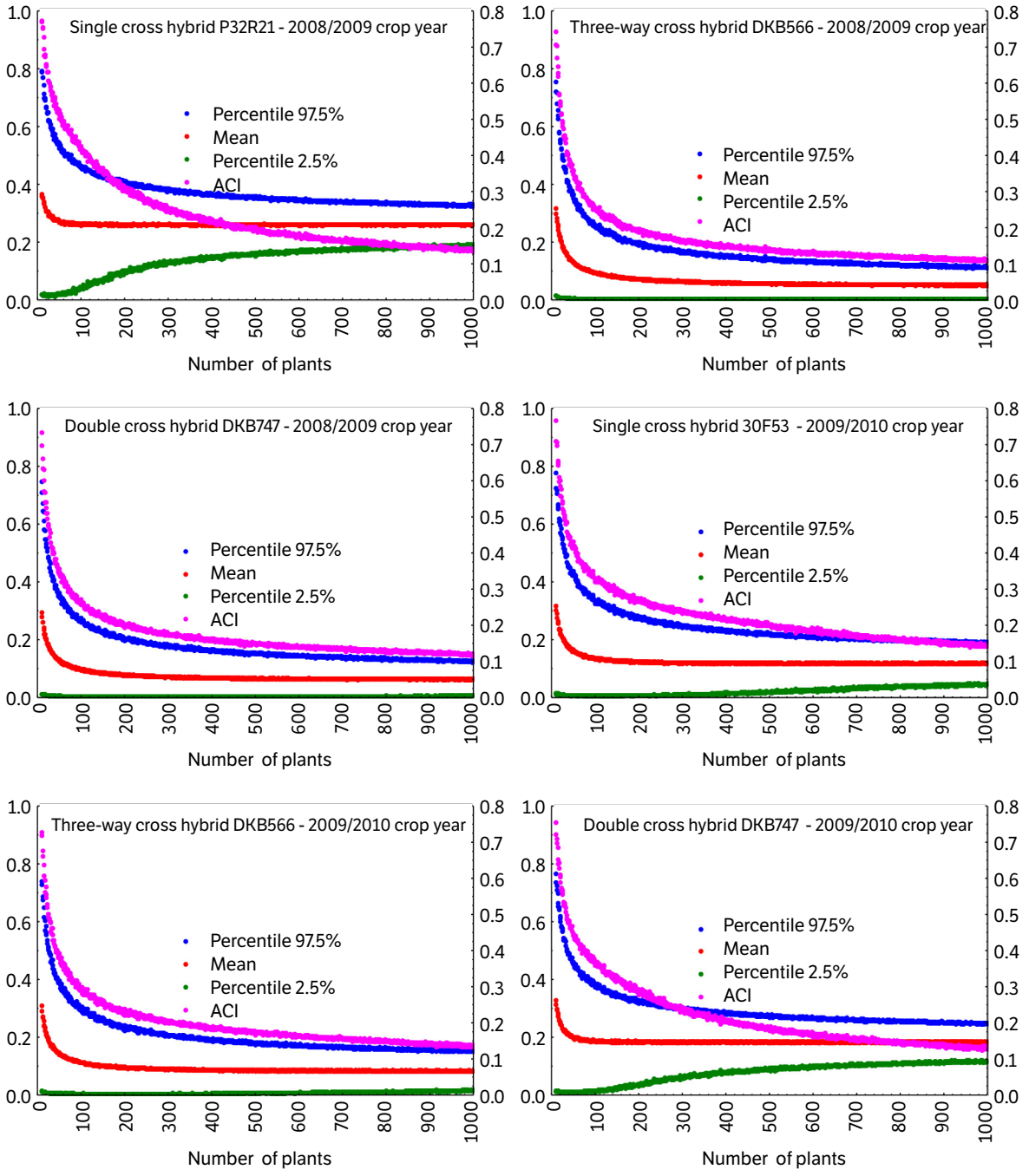

Figure 4. Percentile 97.5%, mean and percentile 2.5% (on the left Y-axis) and amplitude of confidence interval of 95% (ACI) (on the right Y-axis) for 3,000 estimates of second canonical correlation between characteristics of the dimensions of ear (ear length, and ear diameter), and grain production (hundred grains mass, and grains mass per plant) of corn hybrids (*Zea mays* L.), grown in two crop years. On the X axis the number of plants ranges from 10 to 1,000.

DISCUSSION

The condition number in all evaluated cases was less than or equal to 7.44. In this sense, correlation matrices with condition number ≤ 100 are in the class of weak multicollinearity (Montgomery et al. 2012), being possible to carry out

the CCA properly. Both in the first and in the second scenario, the first canonical pair in all cases was significant, being of greater magnitude in the second scenario (Table 1). The second canonical pair had low scores and statistical significance in only a few cases. In this sense, it can be highlighted that statistical significance was obtained due to the high number of observations ($n \geq 361$ plants), and the practical significance is negligible due to the low magnitude (Hair et al. 2009). According to Uurtio et al. (2017), in general, the value of the canonical correlation and the statistical significance are considered jointly to convey the importance of the interrelationships pattern.

Table 3. Estimates of parameters of the model linear response with plateau (a, b), determination coefficient (r^2), sample size (n_o, number of plants) required to estimate the first (CC1) and second (CC2) canonical correlation, and amplitude of confidence interval of 95% in sample size ACI(n_o), between characteristics of the plant architecture (plant height at harvest, and ear insertion height), dimensions of ear (ear length, and ear diameter), and grain production (hundred grains mass, and grains mass per plant) of corn hybrids (*Zea mays* L.), grown in two crop year.

Characteristics	Correlation	a	b	r^2	n_o	ACI(n_o)
Single cross hybrid P32R21 (n = 361 plants) in the 2008/2009 crop year						
Architecture *vs*. grains	CC1	0.51070	-0.00109	0.910	325	0.156
Architecture *vs*. grains	CC2	0.38041	-0.00131	0.853	205	0.111
Ear *vs*. grains	CC1	0.11653	-0.00042	0.860	211	0.027
Ear *vs*. grains	CC2	0.56364	-0.00112	0.915	348	0.173
Three-way cross hybrid DKB566 (n = 373 plants) in the 2008/2009 crop year						
Architecture *vs*. grains	CC1	0.48426	-0.00099	0.901	332	0.154
Architecture *vs*. grains	CC2	0.34106	-0.00042	0.864	486	0.138
Ear *vs*. grains	CC1	0.13183	-0.00047	0.857	215	0.032
Ear *vs*. grains	CC2	0.51376	-0.00229	0.849	164	0.139
Double cross hybrid DKB747 (n = 416 plants) in the 2008/2009 crop year						
Architecture *vs*. grains	CC1	0.49703	-0.00121	0.904	295	0.141
Architecture *vs*. grains	CC2	0.39674	-0.00129	0.864	230	0.100
Ear *vs*. grains	CC1	0.13154	-0.00060	0.844	174	0.027
Ear *vs*. grains	CC2	0.48876	-0.00194	0.843	176	0.148
Single cross hybrid 30F53 (n=1,777 plants) in the 2009/2010 crop year						
Architecture *vs*. grains	CC1	0.42172	-0.00116	0.891	266	0.114
Architecture *vs*. grains	CC2	0.41754	-0.00117	0.824	229	0.150
Ear *vs*. grains	CC1	0.09173	-0.00030	0.877	231	0.023
Ear *vs*. grains	CC2	0.47266	-0.00094	0.823	309	0.181
Three-way cross hybrid DKB566 (n = 1,693 plants) in the 2009/2010 crop year						
Architecture *vs*. grains	CC1	0.46799	-0.00100	0.899	321	0.146
Architecture *vs*. grains	CC2	0.36754	-0.00054	0.818	371	0.167
Ear *vs*. grains	CC1	0.11281	-0.00038	0.861	226	0.028
Ear *vs*. grains	CC2	0.49731	-0.00161	0.841	203	0.169

continue...

Table 3. Continuation...

Characteristics	Correlation	a	b	r^2	n_o	ACI(n_o)
Double cross hybrid DKB747 (n = 1,720 plants) in the 2009/2010 crop year						
Architecture *vs*. grains	CC1	0.47434	-0.00099	0.900	329	0.147
Architecture *vs*. grains	CC2	0.34475	-0.00108	0.845	226	0.101
Ear *vs*. grains	CC1	0.17481	-0.00062	0.862	216	0.042
Ear *vs*. grains	CC2	0.48840	-0.00087	0.891	378	0.161
Overall mean						
Architecture *vs*. grains	CC1	-*	-	0.901	311	0.143
Architecture *vs*. grains	CC2	-	-	0.845	291	0.128
Ear *vs*. grains	CC1	-	-	0.860	212	0.030
Ear *vs*. grains	CC2	-	-	0.860	263	0.162

*Overall mean not calculated.

In a CCA study, Ceccon et al. (2016) found that primary and secondary yield components of maize grains are not independent, being verified canonical correlations of 0.799 and 0.680 for the first and second canonical pair, respectively. According to these authors, plants with higher height, stem diameter, dry mass, and lower ear height positively influence the primary yield components, i.e., dry ear mass, ear length and hundred-grain mass. In a research conducted by Alves et al. (2016) with 27 phenological, morphological, productive, protein-nutritional and energetic-nutritional characteristics in 18 genotypes and three replications (54 plots), the existence of linear dependence between phenological and energetic-nutritional characteristics was verified, and, according to the authors, only phenological characteristics can be used for indirect selection as an indicative of energetic-nutritional quality in corn grains. The authors evaluated the canonical correlation between six pairs of groups and identified canonical correlations of 0.533 to 0.954 for the first canonical pair and from 0.295 to 0.661 for the second canonical pair.

Evaluating 76 corn genotypes in three replications (228 plots) and 29 agronomic, protein-nutritional, and energetic-nutritional characteristics, Alves et al. (2017) verified that the performance of the CCA in the presence of multicollinearity overestimates the variability of canonical coefficients and that the elimination of characteristics is efficient to circumvent the multicollinearity in the CCA. The authors identified canonical correlations of 0.66 to 1 for the first canonical pair and from 0.58 to 1 for the second canonical pair in two scenarios (agronomic characteristics *versus* protein-nutritional characteristics, and agronomic characteristics *versus* energetic-nutritional characteristics), three data bases (36 early maturing corn genotypes, 22 super-early maturing corn genotypes, and 18 transgenic corn genotypes), and two situations (in the presence of multicollinearity or with the elimination of characteristics). In their turn, Crevelari et al. (2019) used CCA to study the linear dependence between groups of morphoagronomic and bromatological characteristics in silage corn hybrids and verified linear dependence between the groups, with the green mass yield associated with crude protein, neutral detergent fiber, lignin, crude fat, and mineral matter. According to the authors, the canonical correlations of the first and second canonical pair were 0.98 and 0.87, respectively, and only the first one was significant.

Based on results showed in Table 2 and Figs. 1, 2, 3 and 4, it can be inferred that CCA generated from a small number of plants should not be considered (less accurate and biased CCA), and that it is important and necessary to define the reference sample size for the generation of accurate CCA. The ACI of CC1 and CC2 gradually decreased with the increase in the number of plants. This result was expected and indicates that the increase in the number of plants provides an improvement in the accuracy of estimates and, consequently, more reliable CCAs. In this sense, Helmer et al. (2021) highlighted that to achieve the stability of the CCA coefficients, larger sample sizes than those normally used are needed. According to the

authors, their results suggest that many studies with CCA might have unstable correlations weights due to an insufficient sample size. Also, according to Helmer et al. (2021), at least 50 samples per variable should be used, and, in cases of SSS, estimated association strengths were too high, and estimated weights could be unreliable for interpretation.

It was found in this study that in small sample sizes the mean of resampling for canonical correlations presented values higher than the original ones (Tables 1 and 2, and Figs. 1, 2, 3 and 4). As the simulated sample size increased, the mean of the canonical correlations decreased and subsequently stabilized. In this sense, Wu and Li (2021) point out that, in cases in which a CCA is obtained from a data set with the dimension larger than the number of samples, it is verified an over-fitting problem arising from the small sample size. According to Helmer et al. (2021), CCA and partial least squares could be highly unreliable when the samples number per feature is relatively small. According to the authors, sample sizes typically used in CCA generate unstable and over-fitting estimates. Also, according to Helmer et al. (2021), stability of models is essential for their replicability, generalizability, and interpretability.

Canonical correlation analyses from small samples size (in this study, less than 270 plants, i.e., subjects per variable ratio smaller that 67.5:1) should be avoided due to inaccuracy of estimates and, therefore, analyses from larger samples (in this study, equal to or greater than 270 plants) should be encouraged. However, after a certain sample size (number of plants) the gains are negligible in relation to the cost for measuring the characteristics of the plants. In this sense, Barcikowski and Stevens (1975) developed a simulation study to evaluate the stability of canonical correlations, canonical weights, and canonical variate-variable correlations and found that the number of subjects per variable to achieve reliability in detecting the most important variables, using components or coefficients, ranged from 42/1 to 68/1. Leach and Henson (2014) verified that bias generally decreased, and the precision of results increased as the sample size to variable ratio increased, with less bias for the 10:1, 25:1, and 40:1 ratio. On the other hand, according to Thompsom (1990), if sample size is at least 10 subjects per variable, canonical results are not as positively biased as some researchers have believed.

In the present study, the lowest ACI was verified for the CC1 estimation of the second scenario (Table 2 for n = 10 and n = 1,000 and Figs. 1, 2, 3 and 4 for the other simulated sample sizes), being that this canonical pair was the one with the greatest association among all the cases studied. These results indicate that the oscillation is smaller in high association canonical pairs. In that regard, Helmer et al. (2021) verified that, as the sample size increases, the strength of the association decreases, reaching the true value faster in CCA with higher correlation coefficients. The authors also found that, for large sample sizes (especially in subjects per variable ratios greater than 100:1), the strength of association assessed in the CCA is very close to the true value, while for small sample sizes the correlations generated from different scenarios are very similar to each other, regardless of the true correlation. Also, according to Helmer et al. (2021), the sample size increases with the number of variables and decreases with the increase of the correlation's strength. The authors proposed a sample sizing calculator and found a pattern of power-law dependence with a strong increase in the required sample size with reduction in the strength between-set correlation.

Other alternatives were pointed out to overcome the problem of insufficient sample sizing in CCA. In that regard, Sun et al. (2010) proposed two-dimensional CCA and concluded that the SSS problem can be effectively solved. Song et al. (2016) realized a combined principal component analysis polymerase chain reaction-CCA approach. According to Uurtio et al. (2017), methods of regularization and the Bayesian CCA are also alternatives to perform CCA in cases of SSS. Krzyśko et al. (2018) demonstrated the use of CCA with multivariate repeated measures, as an interesting alternative in cases of data with small number of observations. Finally, Wu and Li (2021) proposed two alternative methods considered superior to solve these problems in CCA, i.e., the exponential CCA and the randomized exponential CCA.

CONCLUSION

Measuring 270 plants is sufficient to estimate the canonical correlation between groups with two characteristics in each group for corn.

This size can be used as reference for reliable canonical correlation analysis.

AUTHORS' CONTRIBUTION

Conceptualization: Cargnelutti Filho, A. and Toebe M.; **Data curation:** Cargnelutti Filho, A. and Toebe M.; **Formal analysis:** Cargnelutti Filho, A. and Toebe M.; **Investigation:** Cargnelutti Filho, A. and Toebe M.; **Methodology:** Cargnelutti Filho, A. and Toebe M.; **Resources:** Cargnelutti Filho, A. and Toebe M.; **Software:** Cargnelutti Filho, A. and Toebe M.; **Supervision:** Cargnelutti Filho, A. and Toebe M.; **Visualization:** Cargnelutti Filho, A. and Toebe M.; **Writing – original draft:** Cargnelutti Filho, A. and Toebe M.; **Writing – review and editing:** Cargnelutti Filho, A. and Toebe M.

DATA AVAILABILITY STATEMENT

All dataset were generated and analyzed in the current study.

FUNDING

Conselho Nacional de Desenvolvimento Científico e Tecnológico
https://doi.org/10.13039/501100003593
Grant No. 304652/2017-2 and No. 313827/2021-4

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
https://doi.org/10.13039/501100002322
Finance Code 001

ACKNOWLEDGMENTS

To those ones who assisted in carrying out the experiment and in data collection.

REFERENCES

Alves, B. M., Cargnelutti Filho, A., Burin, C. and Toebe, M. (2016). Correlações canônicas entre caracteres agronômicos e nutricionais proteicos e energéticos em genótipos de milho. Revista Brasileira de Milho e Sorgo, 15, 171-185. https://doi.org/10.18512/1980-6477/rbms.v15n2p171-185

Alves, B. M., Cargnelutti Filho, A. and Burin, C. (2017). Multicollinearity in canonical correlation analysis in maize. Genetics and Molecular Research, 16, 1-14. https://doi.org/10.4238/gmr16019546

Barcikowski, R. B. and Stevens, J. P. (1975). A Monte Carlo study of the stability of canonical correlations, canonical weights and canonical variate-variable correlations. Multivariate Behavioral Research, 10, 353-364. https://doi.org/10.1207/s15327906mbr1003_8

Ceccon, G., Santos, A., Teodoro, P. E. and Silva Junior, C. A. (2016). Relationships between primary and secondary yield components of a maize population after 13 stratified mass selection cycles. Journal of Agronomy, 15, 33-38. https://doi.org/10.3923/ja.2016.33.38

Crevelari, J. A., Durães, N. N. L., Santos, P. R., Azevedo, F. H. V., Bendia, L. C. R., Preisigke, S. C., Gonçalves, G. M. B., Ferreira Junior, J. A. and Pereira, M. G. (2019). Canonical correlation for morphoagronomic and bromatological traits in silage corn genotypes. Bragantia, 78, 337-349. https://doi.org/10.1590/1678-4499.20180146

Hair, J. F., Blanck, W. C., Babin, B. J., Anderson, R. E and Tathan, R. L. (2009). Análise multivariada de dados. 6ª ed. Porto Alegre: Bookman.

Helmer, M., Warrington, S., Mohammadi-Nejad, A., Ji, J. L., Howell, A., Rosand, B., Anticevic, A., Sotiropoulos. S. N. and Murray, J. D. (2021). On stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. bioRxiv, 1-73. https://doi.org/10.1101/2020.08.25.265546

Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28, 321-377. https://doi.org/10.2307/2333955

Krzyśko, M., Lukaszonek, W. and Wołyński, W. (2018). Canonical correlation analysis in the case of multivariate repeated measures data. Statistics in Transition New Series, 19, 75-85. https://doi.org/10.21307/stattrans-2018-005

Leach, L. F. and Henson, R. K. (2014). Bias and precision of the squared canonical correlation coefficient under nonnormal data condition. Journal of Modern Applied Statistical Methods, 13, 110-139. https://doi.org/10.22237/jmasm/1398917220

Montgomery, D. C., Peck, E. A. and Vinning, G. G. (2012). Introduction to linear regression analysis. New York: John Wiley and Sons.

Paranaíba, P. F., Ferreira, D. F and Morais, A. R. (2009). Tamanho ótimo de parcelas experimentais: proposição de métodos de estimação. Revista Brasileira de Biometria, 27, 255-268.

R Core Team. (2021). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org

Song, Y., Schreier, P. J., Ramírez, D. and Hasija, T. (2016). Canonical correlation analysis of high-dimensional data with very small sample support. Signal Process, 128, 449-458. https://doi.org/10.1016/j.sigpro.2016.05.020

Sun, N., Ji, Z. H., Zou, C. R. and Zhao, L. (2010). Two-dimensional canonical correlation analysis and its application in small sample size face recognition. Neural Computing and Applications, 19, 377-382. https://doi.org/10.1007/s00521-009-0291-x

Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: a Monte Carlo Study. Educational and Psychological Measurement, 50, 15-31. https://doi.org/10.1177/0013164490501003

Uurtio, V., Monteiro, J. M., Kandola, J., Shawe-Taylor, J., Fernandez-Reyes, D. and Rousu, J. (2017). A tutorial on canonical correlation methods. ACM Computing Surveys, 50, 1-33. https://doi.org/10.1145/3136624

Wu, G. and Li, F. (2021). A randomized exponential canonical correlation analysis method for data analysis and dimensionality reduction. Applied Numerical Mathematics, 164, 101-124. https://doi.org/10.1016/j.apnum.2020.09.013