ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 09, pp.37787-37795, September, 2016

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

REVIEW ARTICLE

PEARSON CORRELATION COEFFICIENTS AND ACCURACY OF PATH ANALYSIS USED IN MAIZE BREEDING: A CRITICAL REVIEW

***,[1]Tiago Olivoto, [2]Maicon Nardino, [3]Ivan Ricardo Carvalho, [4]Diego Nicolau Follmann, [5]Vinícius Jardel Szareski, [3]Mauricio Ferrari, [3] Alan Junior de Pelegrin and [6]Velci Queiróz de Souza**

[1]Department of Agronomic and Environmental Sciences, Federal University of Santa Maria Frederico Westphalen, Rio Grande do Sul, Brazil
[2]Department of Mathematics and Statistics, Federal University of Pelotas, Capão do Leão, Rio Grande do Sul, Brazil
[3]Plant Genomics and Breeding Center, Federal University of Pelotas, Capão do Leão, Rio Grande do Sul, Brazil
[4]Agronomy Department, Federal University of Santa Maria, Santa Maria, Rio Grande do Sul, Brazil
[5]Dept. of Crop Science, Federal University of Pelotas, Capão do Leão, Rio Grande do Sul, Brazil
[6]Federal University of Pampa, Dom Pedrito, Rio Grande do Sul, Brazil

ARTICLE INFO

Article History:
Received 17th June, 2016
Received in revised form
16th July, 2016
Accepted 29th August, 2016
Published online 20th September, 2016

Key words:
Zea mays. Average data.
Correlation matrices,
Systematic errors.

ABSTRACT

Maize (*Zea mays* L.) has been the subject of several studies involving correlation coefficient estimates and path analysis. This critical review discusses some systematic errors that have been observed in estimating of correlation coefficients and its possible impacts on accuracy of path analysis. In a first moment, an approach about the maize crop, origin, characteristics and biometric models commonly used in genetic breeding of this crop is presented. Some obstacles found in estimates of path coefficients and the methods used to adjust them are discussed. We also present evidences and a theoretical explanation that some data arrangement methods currently used, may be overestimating the correlation coefficients in scientific studies. Data from a literature search revealing the accuracy of path analysis of some research are presented and discussed. In a last moment, we present a future perspective about how the correct estimate of the correlation coefficients may improve the accuracy of path analysis, underscoring the need for research directed to this objective.

Citation: ***Tiago Olivoto, Maicon Nardino, Ivan Ricardo Carvalho et al.*****2016.** "Pearson correlation coefficients and accuracy of path analysis used in maize breeding: a critical review", *International Journal of Current Research*, 8, (09), 37787-37795.

INTRODUCTION

Maize crop

Maize (*Zea mays* L.), belonging to the grass family Poaceae is the most produced cereal at the world, surpassing the mark of 1 billion tons produced in 2014 growing season. This crop has great economic, social and recently environmental importance due their grain serve as alternative raw material for ethanol production (Hertel *et al.,* 2010).

Corresponding author: Tiago Olivoto
Department of Agronomic and Environmental Sciences, Federal University of Santa Maria Frederico Westphalen, Rio Grande do Sul, Brazil.

The world's leading producers of this cereal is the United States, China and Brazil. It is known that the currently known maize is the result of a long evolutionary process, being the hypothesis most accepted that it evolved from the teosinte, with the center of origin in Central America, specifically in Mexico. During this process, genetic events were decisive for the change in plant architecture and crops' inflorescence characteristics. Two quantitative traits loci (QTL) were identified as the main responsible for the morphological differences between these species. The first (TB1) located on arm of chromosome 1L has effects on the gender of the inflorescence and the number and length of internodes on the lateral branches; the second, located on arm of chromosome

3L, affects the same characteristics. A study evaluating the segregation of these loci revealed that they present epistatic interrelationships turning together, substantially, the plant architecture and inflorescence (Doebley *et al.,* 1995). The number of chromosomes present in the modern maize is 10, but it has long been suspected that this number was the result of a historical tetraploid event. Several observations point to this possibility, including the fact that the culture have duplicated chromosome segments (Gaut, 2001). Some of these segments were sequenced and the standard divergence between 14 pairs of duplicated genes was examined. The results indicated that the time in this sequences' duplication vary in two distinct groups, corresponding to about 20.5, and 11.4 million years ago. This observation indicates the possibility of a allotetraploid genomic event where his two diploid progenitors diverged about 20.5 million years ago, and that the allotetraploid event probably occurred approximately 11.4 million years ago (Gaut and Doebley 1997).

Maize breeding

It is attributed to Darwin the first works with plant selfing, however, were East and Shull the pioneers in the study of the influence of successive selfing and exploitation of heterosis in maize. During the era of hybrid maize (1908 to present), the crop yield has increased almost six times (Lee and Tollenaar, 2007). In early 1908, George Harrison Shull, published a paper with the title 'The composition of the field of maize', marking the beginning of the exploitation of heterosis in plant breeding, certainly one of the greatest genetic triumphs of our time. In his work, Shull showed that inbred lines of maize, subjected to several cycles of selfing showed significant reduction in vigor and grain yield; however, the hybrids resulting from two inbred lines had these features recovered, often featuring performance and superior vigor of varieties from which the inbred lines were derived (Shull, 1908). At the same time, Edward Murray East, made similar experiments and also recognized the deleterious effects of inbreeding in maize plants; however, did not realize the value of crossing inbred lines, up to study Shull's paper. East was not convinced of the usefulness of the idea, because, really, inbred lines produced a very small amount of seeds, burdening any increase in production provided by hybrids. Both were at odds, but have remained true to their findings (Crow, 1998)

The limitation in seeds' production was surpassed later (1918) from an idea of Donald Forsha Jones, who while still a graduate student, defended the idea of using four genetic bases, or double-cross hybrids. The principle involved crossing two inbred lines and later, crossing of this hybrid with another, resulting from two other inbred lines. These hybrids were somewhat more variable compared with simple hybrids, however, much less than the open-pollinated varieties existing at that time. As seeds were coming from a simple hybrid, the largest quantity of available seed improved the program viability (Jones, 1918). Increases in maize productivity was, no doubt, largely due to the discovery of heterotic effect; however, the evolution in agricultural practices, such as increased use of fertilizers, changes in plant's arrangement, cultivation practices and agricultural mechanization, were useful tools and that combined with the use of higher-genetically plants enabled the achievement of high yields currently observed. But, it would be possible to separate the contribution of these effects? Studies evaluating the productivity of maize in a period of 70 yr. showed an average increase of 65-75 kg ha^{-1} yr^{-1}, and that genetic breeding was responsible for about 50% of this increase (Duvick, 2005, 1977). A maize ideotype had been proposed by (Mock and Pearce, 1975). The ideotype that should produce optimally when grown in an environment without limitations of edaphoclimatic factors, high plant density and reduced spacing between rows, it is characterized by: a) rigid vertically-oriented leaves above ear (leaves below the ear should be horizontally-oriented); b) maximum photosynthetic efficiency; c) efficient conversion of assimilates in grains; d) short interval between pollination and the emergence of style-stigmas; e) prolificacy; f) small size of cobs; g) insensitivity to photoperiod; h) cold tolerance in the germination (for cultivated genotypes in areas where early sowing takes place in cold or wet soil); i) as long as possible grain filling; and j) slow leaves senescence.

In this regard, studies aiming at a higher-plant architecture (Tian *et al.,* 2011), better floral sync (Buckler *et al.,* 2009), improved photosynthetic efficiency (Fracheboud *et al.,*1999) and absorption of nutrients (Gallais and Hirel, 2004) has been successful. The combination of all the favorable characteristics in a single hybrid, however, is a daunting task for breeders mainly due, in most part, the traits are expressed by different genic actions (Sa *et al.,* 2014). Success in maize breeding, as well as in other economically important crops also was due to wider use of statistic-experimental models in the selection of superior hybrid, introduced by Fisher, involving repetition, randomization and local control. The author states the importance of a thorough selection in a plant breeding program. In the case of simple maize hybrids in particular, this process occurs in three steps. 1) choice of individuals in a population to start the process; 2) artificial selfing of these individuals aiming to inbreeding and selection of pure lines and 3) artificial crosses. If plants are randomly selected in each step, the hybrids are actually a random sample of the original population. Thus, the criteria-based selection in the three steps should be considered. At first, the selection resembles the mass selection, practiced in breeding of open-pollinated varieties. In the second, the selection is neutralized quickly by rapid fixation, due to homozygosity increase in 50% each generation; so Fisher emphasized that the selection in the last step, should be greater emphasis. In fact, the selection at this step is important as it is being practiced in the studied subject (Fisher, 1925).

Biometric models used in maize hybrids

Several statistical models has been use to evaluate the performance of maize hybrids. Models that allow the partition of genotype x environment interaction into environmental and genetic components are useful to evaluate the adaptability and stability of hybrids, especially in assessment of value for cultivation and use. Mixed models with fixed and random variance components has also proved efficient to identify promising hybrids in breeding programs (Baretta *et al.,* 2016). Knowledge of association degree between traits is of fundamental importance in plant breeding programs. This importance increases, especially if some desirable trait present

difficulty in assessment, or low heritability (Cruz *et al.,* 2014). The Pearson product-moment correlation coefficient (Pearson, 1920), has been widely used for this purpose. Although this correlation reveals the direction and degree of linear association between a pair of traits, it does not show interrelationships of cause and effect. Thus, Sewall Wright in his work entitled 'Correlation and causation' (Wright, 1921) proposed a method known as 'path analysis' allowing this understanding. The method is based on partitioning of the linear correlation coefficient into direct and indirect effects of a group of explanatory traits on response of a dependent trait. Path analysis has been highlighted in breeding area because the selection aiming improving a desirable trait that has difficulty-measure and low heritability, can be indirectly carried out by another trait, directly associated with desirable trait, but that shows high heritability and is easy to measure. In maize, as well as in several world-important crops, studies using path analysis has been successful in the sense of revealing the interrelationships between traits, be them yielding, grain quality or the effects of interaction genotype x environment or management of cultivation (Adesoji *et al*, 2015, 2015; Jadhav *et al*, 2014; Ma *et al*, 2015; Nardino *et al*, 2016). Studies with path analysis in maize were succeeded in revealing the interrelationships between yield components. In summary, the results converge to a common conclusion: the number of kernels per ear and thousand-kernel weight are the traitswith greater direct association with grain yield (Adesoji *et al.,* 2015; Khameneh *et al.,* 2012; Mohammadi *et al.,* 2003; Reddy *et al.,* 2012). As the heritability in the broad sense of these trait is high (> 0.90), indirect selection from these traits aiming at increasing grain yield (trait highly influenced by the environment) can be effective (Ojo *et al.,* 2006).

Path analysis conception

Path analysis is originally based on ideas developed by Sewall Wright (Wright, 1921), however from its conception to the method's consolidation, some disagreement about the reliability of the mathematical method that explains the relationships of cause and effect were observed. In 1922, Henry E. Niles, in his paper entitled "Correlation, Causation and Wright's theory of path coefficients", made a criticism of the method proposed by Wright, claiming that the philosophical basis of the path coefficients method was doubtful. Niles, testing Wright's method, had observed in some of its results correlations exceeding | 1 |, saying "these results are ridiculous" and that Wright would have to provide much more convincing evidences than he was presenting (Niles, 1922). In the following year, Sewall Wright in his paper entitled "The theory of path coefficients: a reply to Niles's criticism", consolidates his method concluding that Niles seemed to be based on incorrect mathematical concepts, result of a failure to recognize that path coefficient it is not a symmetric function of two traits, but it necessarily has direction. Wright concludes his work by stating that the path analysis does not provide a formula to infer causal relationships from knowledge of the correlations; it is, however, within certain limitations, a method of evaluating the logical consequences of a causal hypothesis relationship in a system of correlated traits. It adds that the criticism offered by Niles nothing invalidates the theory or application of path coefficient (Wright, 1923). Currently, the statistical method of path coefficient is consolidated and worldwide used in several areas of science. In order to estimate path coefficients, normal equations models are used to partition the linear coefficients into direct and indirect effects of a set of explanatory traits on a dependent trait. Thus, their estimates need a previously-estimated linear correlation matrix among traits in study.

Estimation of linear correlation

One of the most used measures in breeding in order to estimate the direction and degree of linear association between two random traits is the Pearson product-moment correlation coefficient. To estimate the degree of association between two hypothetical traits X and Y, let's consider the following assumption. The traits should form the following dataset. (X_1, Y_1), $(X_2, Y_2)...(X_n, Y_n)$. Thus, correlation coefficient estimates between X and Y is obtained by the following equation:

$$r = \frac{\sum_{i=1}^{n}\{(X_i-\bar{X})(Y_i-\bar{Y})\}}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

Were, $\sum_{i=1}^{n}\{(X_i-\bar{X})(Y_i-\bar{Y})\}$ is the covariance XY; $\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$ is the product of standard deviation of Xand Y; $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ e $\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$.

Although the merit of this analysis had been attributed to Karl Pearson, the method was originally designed by Francis Galton, who defined the term correlation as to the following: "two variables are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction" (Galton, 1888). So, your estimate takes into account the covariance between two traits, represented here by XY divided by the product of respective standard deviation of X and Y. Taking into account the premise of this analysis, the traits which will be correlated, have being, mandatorily, assessed in the same subject, in order to represent the actual covariance and standard deviation of the set of observations.

Path analysis estimation

After obtaining linear correlation estimates (r), partitioning of linear correlations into direct and indirect effects of an explanatory dataset with p-traits can be performed by derivation of the set of normal equations ($X'X\beta = X'Y$) in order to estimate parameters of multiple regression using OLS. Thus, β estimate is given by: $\beta = X'X^{-1} X'Y$, where β is the partial regression coefficient ($\beta_1, \beta_2, \beta_3,...,\beta_p$) to p + 1 rows; $X'X^{-1}$is the inverse of linear correlation matrix among explanatory traits; and X'Y is the correlation matrix between each explanatory trait with the dependent trait. After estimating the regression coefficients (β_p), the direct and indirect effects of a set of p-explanatory trait towards the dependent trait can be estimated.

Consider the following example, where a set of explanatory traits (a, b, c and d) are used to explain the relationship of cause and effect on the response of dependent variable (y). After partial regression estimations (β_1, β_2, β_3 eβ_4), direct and indirect effects of 'a' on 'y' are given by: $r_{a:y} = \beta_1 + \beta_2 r_{a:b} + \beta_3 r_{a:c} + \beta_4 r_{a:d}$, where $r_{a:y}$ is the linear correlation between 'a' e 'y', β_1 is the direct effect of 'a' on 'y'; $\beta_{2a:b}$is the indirect effect of 'a' on 'y' via 'b', $\beta_{3a:c}$ is the indirect effect of 'a' on 'y' via 'c' and$\beta_{4a:d}$is the indirect effect of 'a' on 'y' via 'd'. Similar equations are used in order to estimate direct and indirect effects of b, c, and d. The coefficient of determination of the model, i.e., how much of the variance in the dependent trait is explained by the interrelationship on explanatory traits, is given by $R^2 = \beta_1 r_{a:y} + \beta_2 r_{b:y} + \beta_3 r_{c:y} + \beta_4 r_{c:y}$. Residual effect is estimated by: Noise = $\sqrt{1 - R^2}$.

This technique has facilitated the understanding of the interrelationship among traits and their effects on dependent trait in several areas of science, as in plant breeding and crop management (Abdala *et al.,* 2016; Dewey and Lu, 1959; Farooq *et al.,* 2015; Mohammadi *et al.,* 2016; Nardino *et al.,* 2016; Olivoto *et al.,* 2015; Souza *et al.,* 2015), animal breeding (Norris *et al.,* 2015; Önder and Abaci, 2015), environmental and social sciences (Hong *et al.,* 2016; Xu *et al.,* 2014), humanities (Hagger *et al.,* 2016) and several related areas. Indeed, path analysis has been a useful tool particularly in plant breeding, however, care must be taken prior the estimation of this analysis. Below we discuss some obstacles encountered in the estimates of the path coefficients.

Difficulties observed in path analysis

Although this analysis shows associations of cause and effect, its estimate is based on multiple regression principles. thus, it can be biased by complex nature of the data, wherein the response of the dependent trait is linked to a large number of explanatory traits that are often correlated between them (Graham, 2003). Correlated traits are difficult to analyze because its effect on the response variable may be due to any synergistic relationship between variables or spurious correlations. Thus, where two explanatory traits are highly associated, it is difficult to estimate the relationship of each individual explanatory trait, since these, as a whole contribute to the explanation of the linear relationship. This particularity is known as multicollinearity (Blalock, 1963).

Matrices multicollinearity

What is it?

In multiple linear regression, data is fitted to a multiple linear model that predicts the values of a response variable (Y) from the weighted sum of several explanatory traits (Xi) and the random error (ε). $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i + \varepsilon$. Where β are regression coefficients. The main goal is to fit a model using the smallest number of traits that explain the most variance of response variable. If all explanatory traits are independent, each of the regression coefficients (βi) represent the total contribution of a given predictor in response trait*;* if, however, two or more explanatory traits are associated, partial regression coefficients need to be estimated to isolate the contribution of a single explanatory trait. The distinction between single contributions is the crucial point in multiple regression analysis and also the largest inferential problem encountered due to the presence of multicollinearity (Graham, 2003; Gunst and Mason, 1977). When this phenomenon occurs in moderate or severe levels, the variances associated with path estimators can reach too high values, making unreliable estimates. Montgomery *et al* (2015), proposed a classification for multicollinearity based on the condition number (CN), i.e. the ratio between the largest and smallest eigenvalue of explanatory traits matrix. Thus, the degree of multicollinearity is considered weak, moderate and severe when CN ≤100 between 100 and 1000 and ≥ 1000, respectively. Another indicator used to identify the presence of multicollinearity is called variance inflation factor (VIF), which as the name suggests, quantifies how much the variance of regression coefficient is inflated. For each of β coefficients in a multiple regression model, there is one VIF. When the VIF for a given predictor is 1, it means that there is no correlation between the predictor and the remainder of predictor traits. This fact is hardly observed. Can be taking as a rule, that the existence of VIFs greater than 10, are serious multicollinearity signals, being necessary to take some action to adjust it (Mansfield and Helms, 1982; O'Brien, 2007). Path coefficients at odds with biological expectation were observed when the analysis was performed in the presence of severe multicollinearity (Toebe and Cargnelutti, 2013). In addition to this, a study by (Petraitis *et al.,* 1996) revealed that from 24 path analysis published in ecological studies, 15 had problems with multicollinearity, resulting in 13 cases with biased path coefficients. This information is worrying because in the case of plant breeding, path coefficients wrongly estimated and interpreted, may result in an inefficient selection, bringing into play the financial, human and time spent in the conduct of a plant breeding program.

Methods for adjusting multicollinearity

Although the problems related to multicollinearity presents itself as a difficulty in estimating path coefficients, some steps can be taken to mitigate its undesirable effects when it is detected by the aforementioned methods. It is now known that the exclusion of the traits responsible for inflating the variance of a regression coefficient is an effective technique and reduces the multicollinearity in matrices of explanatory traits (Jadhav *et al.,* 2014). The identification of these traits, however, can become a difficult task. As previously discussed, the purpose of multiple regression (path analysis) is to identify a set of explanatory traits with high power, but which do not exhibit highly correlated. In this sense, there are several variable selection methods to choose a subset of predictors with minimal multicollinearity, such as hierarchical models, stepwise procedures and criteria-based models (George and McCulloch, 1993; Mitchell and Beauchamp, 1988; Nishii, 1984; Wold *et al.,* 1984). In a focused approach to plant breeding Cruz *et al*. (2014), discuss a method to identify the traits responsible for multicollinearity in a set of explanatory traits. This method is based on analysis of eigenvalues and eigenvectors of a symmetric positive definite matrix of explanatory traits and identifies the traits responsible for this problem, as that with the highest weight (component of the

eigenvector) associated to the eigenvalues of lesser magnitude. The exclusion of traits responsible for multicollinearity allowed estimating path coefficients, without its harmful effect, in research with several crops such as rice (Shrivastava and Sharma, 1976), canola (Coimbra *et al.,*1999), soybean (Bizeti *et al.,*2004) and maize (Toebe and Cargnelutti, 2013). It should be noted that the choice of traits for exclusion must be carefully, because traits with high explanatory power removed from the model, can reduce the coefficient of determination (R^2), and increase the noise's model (Cross *et al.,*2014).When the exclusion of multicollinearity-generating traits is not a procedure considered by researcher, e.g., due to a small number of explanatory traits, or the importance of knowing their effects, a second option is to perform the path analysis with all the explanatory traits, but with the addition of a small value in diagonal elements of X'X, known as ridge regression (Hoerl and Kennard, 1970).

This method aims to reduce the variance associated with the OLS estimators. Thus, β estimates obtained in ridge regression are obtained similarly to the conventional method, however solving the partially-modified normal equations system $(X'X+k)\beta = X'Y)$ generating $\beta = (X'X+k)^{-1} X'Y$, for $0< k<1$. Where, β is the partial coefficient regression ($\beta_1, \beta_2, \beta_3,...,\beta_p$) to p + 1 rows; $(X'X+k)^{-1}$is the inverse of linear correlation among explanatory traits with k constant included in diagonal elements; and X'Y is the correlation matrix between each explanatory traits with dependent trait. Using numerical examples in order to illustrate the effectiveness of this method, Marquardt (1970), concluded that the ridge regression method is efficient in estimating path analysis coefficients from non-orthogonal data. In plant breeding, this technique also had proved effective in improving the conditioning of explanatory traits matrices in research of several economically-important crops (Bizeti *et al.,* 2004; Coimbra *et al.,* 1999; Luz *et al.,* 2011; Nardino *et al.,* 2016; Nogueira *et al.,* 2012; Olivoto *et al.,* 2015; Souza *et al.,* 2015)

Can multicollinearity be reduced?

Although the techniques for adjusting multicollinearity has been effective and widely known, such techniques are used after the diagnosis of the correlation matrix among explanatory traits, that is, its use is only possible after the estimation of linear correlation matrix. As previously discussed, multicollinearity is directly associated with high magnitude of correlation between explanatory traits in the model. In this sense, in order to estimate the actual correlation between two random traits (X and Y), the covariance and standard deviation should represent the population under study. In agronomic experiments, it is common assessing of several samples (plants) in each plot of each treatment, to represent the population (treatment). Such plants routinely make up an average of this specific plot, which will be used later for ANOVA and supplementary analysis, such as multiple-comparation analysis. In a bibliographic research project, were found, however, several studies that has been using these averages to estimate the correlation coefficients and then the path coefficients (Adesoji *et al.,* 2015; Faria *et al.,* 2015; Khameneh *et al.,* 2012; Kumar and Babu, 2015; Nataraj *et al.,* 2015, 2014; Rigon *et al.,* 2012; Toebe and Cargnelutti, 2013; Torres *et al.,* 2015). Starting from the assumption that the average can mask the individual variances (of assessed plants), correlations estimated from these average do not represents the actual variance and standard deviation of the traits (x, y,... z) of the original population. In addition to the statistical concept methodologically biased, the inference of magnitude and direction of interrelationships between traits when the correlation is estimated with average data is misleading, because this inference is performed in a different population of the original (e.g. when all plants are used for this estimate). As large number of agronomic studies makes populational inferences based on sampling (plants), using the average value of these plants to estimate correlations and make inference to the original population, it is, without doubt, a misconception that should be considered.

A theoretical explanation

We take as an example an experiment to evaluate the direction and degree of association between trait of maize hybrids, conducted in a randomized block design with 5 treatments (simple hybrid) and four replications. In each replication (plot) is common to assess traits in several plants, aiming to represent the population of this specific plot. In experiments with maize hybrids usually are sampled 3 to 5 plants per plot, mainly because they present low phenotypic variation. So in this hypothetical experiment, we assume that in five plants of each plot were evaluated three traits (X, Y, Z). Researcher would then have the values of these three traits assessed in 100 plants (5 hybrids x 4 replications x 5 plants). To estimate the correlation between X and Y, e.g., the following dataset is required: (X_1, Y_1), (X_2, Y_2), ..., (X_{100}, Y_{100}). Correlation coefficient is then given by applying the formulae described in "estimation of linear correlation". When the researcher uses the average values of plots in order to correlation estimating, he is masking the deviations of each trait (X, Y and Z) relative to the overall average of these traits. In this case, the observed deviations among the five plants of each plot will be canceled out by the average of these plants.

The new data set used for the same estimation of the correlation between X and Y in this methodology will then be: (X_1, Y_1), (X_2, Y_2), ..., (X_{20}, Y_{20}). The observed variance in the new dataset is then representing variance of average from five original sampled plants, and not the variance coming from all these plants; therefore, this variance is masked, and tends to present itself lower, compared to the original variance. This fact should be taken into account, because the inference of the direction and magnitude of association between characters is being made for a different population of the original. After in-depth evaluation of the correlation formula 'see estimation of linear correlation', it is noted that the formulae's divisor is estimated by product of the standard deviations of X and Y. Then, when the correlation is estimated based on average data, generally showing less variation, the product of these deviations will present smallest. Assuming that the covariance between X and Y remain similar, dividing by a smallest divider, will result in a coefficient of correlation overestimated. But, could this mistake found in the correlation estimates be associated with higher multicollinearity problems in explanatory traits matrices and with the reduction of accuracy

in path analysis? This approach, as far as we known, is still lacking in the literature.

Path analysis accuracy in ecological experiments

In a randomized research of 25 studies using path analysis, we observed a certain contradiction regarding to information of the coefficient of determination (R^2) and model's noise. For example, only five studies (20%) clearly showed the R^2 and the noise in their results. In four studies (16%), only the R^2 was presented, while in six studies (24%) only the noise was presented. In 10 studies (40%), neither of these parameters were found. This is alarming, because it can mask the interpretation of the reader in not to know how much of the variation in the dependent trait was explained by the model. In studies that showed adjustment measures, were observed R^2 fluctuating between 0.31 and 0.99 and noises ranging from 0.105 to 0.680. It is also observed that in some cases, the noise approached of the R^2, a fact that may cast doubt on the reliability of the estimated path coefficients (Table 1).

Table 1. Multiple Coefficient of Determination (R^2) and the residual effect observed in 25 studies involving path analysis

Species	R^2	Residual	Reference
Castor beans	0.89	np	Torres *et al*., 2015
Cotton	np†	np	Farooq *et al*., 2015
Indigenous goats	np	np	Norris *et al*., 2015
Maize	np	0.345	Adesoji *et al*., 2015
Maize	np	np	Agrama, H. 1996
Maize	np	np	Alvi *et al*., 2003
Maize	np	np	Bello *et al*., 2010
Maize	np	0.560-0.670	Carvalho *et al*., 2001
Maize	0.555	0.667	Faria *et al*., 2015
Maize	0.31-0.99	np	Khameneh *et al*., 2012
Maize	np	0.249	Kumar *et al*., 2011
Maize	np	0.105	Kumar *et al*., 2013
Maize	0.851	0.386	Kumar *et al*., 2015
Maize	np	0.372	Nataraj *et al*., 2014
Maize	np	np	Nataraj *et al*., 2015
Maize	0.64	0.53	Rigon *et al*., 2012
Maize	np	np	Saleem *et al*., 2007
Maize	0.74‡	0.490‡	Toebe and Cargneluti, 2013
Peanut	np	np	Luz *et al*., 2011
Pearl millet x Elephantgrass	np	np	Diz *et al*., 1994
Rice	0.915	np	Abdala *et al*., 2016
Soybean	0.912	0.295	Bárbaro *et al*., 2006
Soybean	0.909-0.950	np	Bizeti *et al*., 2004
Soybean	np	np	Iqbal *et al*., 2004
Wheat	np	0.470-0.680	Khan and Naqvi, 2012

† np, not presented.
‡ Average from 14 path analysis.

Future perspectives

Research aimed to demonstrate if and how much the use of average values may overestimate the correlation coefficients, increase multicollinearity in analysis that use multiple regression and reduce its accuracy are necessary and certainly will be welcomed. Thus, by combining the correct estimate of correlation coefficients with the known methods to adjust multicollinearity, the accuracy of path analysis in biological studies could be improved. It is noteworthy that, completely eliminate the multicollinearity in matrices of explanatory traits is an almost impossible task, because the degree of interrelationship coming from the nature of the traits is inevitable. From a breeding viewpoint, the effectiveness of indirect selection based on path coefficients will depend then of: (i) researcher's ability to correctly estimating correlation coefficients; (ii) take the right steps to adjust multicollinearity of their matrices; (iii) include in group of predictors, traits that explain most of the observed variance in the dependent trait; and (iv) carry out the selection based on traits with high heritability and which are directly associated with the response in dependent trait.

Final Considerations

Path analysis has been helping researchers from several areas of science in order to reveal logical relationships of cause and effect. In maize genetic breeding, in particular, this technique has allowed the knowledge of the interrelationships between traits, enabling faster-indirect selection of lines in inbreeding process. The methods currently used for adjusting the multicollinearity of explanatory traits matrices are effective. Observation of studies with correlation coefficients tendentiously estimated and also studies in which have been hidden important information, such as coefficient of determination and model's noise, however, is worrying. In this sense, research aiming to compare the influence of average values on estimates of correlation coefficients and its impact on path analysis accuracy are needed, and could may help researchers reduce systematic errors in their experiments.

REFERENCES

Abdala, A.J., Bokosi, J.M., Mwangwela, A.M., and Mzengeza, T.R. 2016. Correlation and path co-efficient analysis for grain quality traits in F1 generation of rice (*Oryza sativa* L.). *Journal of Plant Breeding and Crop Science,* 8:109–116.

Adesoji, A.G., Abubakar, I.U., and Labe, D.A. 2015. Character Association and Path Coefficient Analysis of Maize (*Zea mays* L.) Grown under Incorporated Legumes and Nitrogen. *Journal of Agronomy*, 14:158–163.

Agrama, H. A.S. 1996. Sequential path analysis of grain yield and its components in maize. *Plant Breeding,* 115:343–346.

Alvi, M.B., Rafique, M., Tariq, M.S., Hussain, A., Mahmood, T., and Sarwar, M. 2003. Character association and path coefficient analysis of grain yield and yield components maize (Zea mays L.). *Pakistan Journal of Biological Sciences*, 6:136–138.

Bárbaro, I.M., Centurion, M.A.P.D.C., Di Mauro, A.O., Unêda-Trevisoli, S.H., Arriel, N.H.C., and Costa, M.M. 2006. Path analysis and expected response in indirect selection for grain yield in soybean. *Crop Breeding and Applied Biotechnology,* 6:151–159.

Baretta, D., Nardino, M., Carvalho, I.R., Oliveira, A.C. de, Souza, V.Q. de, and Maia, L.C. da, 2016. Performance of maize genotypes of Rio Grande do Sul using mixed models. *Científica,* 44:403-411.

Bello, O.B., Abdulmaliq, S.Y., Afolabi, M.S., and Ige, S.A. 2010. Correlation and path coefficient analysis of yield and agronomic characters among open pollinated maize varieties and their F 1 hybrids in a diallel cross. *African Journal of Biotechnology* 9:2633–2639.

Bizeti, H.S., Carvalho, C.G.P. de, Souza, J.R.P. de, and Destro, D. 2004. Path analysis under multicollinearity in soybean. *Brazilian Archives of Biology and Technology* 47:669–676.

Blalock, H.M., 1963. Correlated Independent Variables: The Problem of Multicollinearity. *Social Forces,* 42:233–237.

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C., Goodman, M.M., Harjes, C., Guill, K., Kroon, D.E., Larsson, S., Lepak, N.K., Li, H., Mitchell, S.E., Pressoir, G., Peiffer, J.A., Rosas, M.O., Rocheford, T.R., Romay, M.C., Romero, S., Salvo, S., Villeda, H.S., Silva, H.S. da, Sun, Q., Tian, F., Upadyayula, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., and McMullen, M.D. 2009. The Genetic Architecture of Maize Flowering Time. *Science* 325:714–718.

Carvalho, C.G.P. de, 2001. Path analysis under multicollinearity in So x So maize hybrids. *Crop Breeding and Applied Biotechnology,* 1:263–269.

Coimbra, J.L.M., Guidolin, A.F., Carvalho, F.I.F., Coimbra, S.M.M., and Marchioro, V.S. 1999. Análise de trilha I: análise do rendimento de grãos e seus componentes. *Ciência Rural,* 29:213–218.

Crow, J.F. 1998. 90 Years Ago: The Beginning of Hybrid Maize. *Genetics* 148:923–928.

Cruz, C.D., Carneiro, P.C.S. and Regazzi, A.J., 2014. Modelos Biométricos Aplicados ao Melhoramento Genético, 3rd ed. UFV, Viçosa, MG.

Dewey, D.R., Lu, K., 1959. A correlation and path-coefficient analysis of components of crested wheatgrass seed production. *Agronomy Journal* 51:515–518.

Diz, D.A., Wofford, D.S., Schank, S.C., 1994. Correlation and path-coefficient analyses of seed-yield components in pearl millet x elephantgrass hybrids. *Theoretical and Applied Genetics,* 89:112–115.

Doebley, J., Stec, A., and Gustus, C. 1995. teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* 141:333–346.

Duvick, D.N. 1977. Genetic rates of gain in hybrid maize yields during the past 40 years. *Maydica* 22:187–196.

Duvick, D.N. 2005. Genetic progress in yield of United States maize (*Zea mays* L.). *Maydica*, 50:193-202.

Faria, L.A., Peluzio, J.M., Afférri, F.S., Carvalho, E.V. de, Dotto, M.A., and Faria, E.A. 2015. Análise de trilha para crescimento e rendimento de genótipos de milho sob diferentes doses nitrogenadas. *Journal of Bioenergy and Food Science,* 2:1–11.

Farooq, J., Anwar, M., Rizwan, M., Riaz, M., Mahmood, K., and Mahpara, S. 2015. Estimation of Correlation and Path Analysis of Various Yield and Related Parameters in Cotton (*Gossypium hirsutum* L.). *Cotton Genomics and Genetics,* 6:1–6.

Fisher, R.A., 1925. Statistical Methods for Research Workers, 1st ed. Oliver and Boyd, London.

Fracheboud, Y., Haldimann, P., Leipner, J., and Stamp, P. 1999. Chlorophyll fluorescence as a selection tool for cold tolerance of photosynthesis in maize (*Zea mays* L.). *Journal of Experimental Botany,* 50:1533–1540.

Gallais, A., and Hirel, B. 2004. An approach to the genetics of nitrogen use efficiency in maize. *Journal of Experimental Botany*, 55:295–306.

Gaut, B.S. 2001. Patterns of Chromosomal Duplication in Maize and Their Implications for Comparative Maps of the Grasses. *Genome Research*, 11:55–66.

Gaut, B.S., and Doebley, J.F. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences*94:6809–6814.

George, E.I., and McCulloch, R.E. 1993. Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88:881–889.

Graham, M.H., 2003. Confronting Multicollinearity in Ecological Multiple Regression. *Ecology*, 84:2809–2815.

Gunst, R.F., and Mason, R.L. 1977. Advantages of examining multicollinearities in regression analysis. *Biometrics,* 33:249–260.

Hagger, M.S., Chan, D.K.C., Protogerou, C., and Chatzisarantis, N.L.D. 2016. Using meta-analytic path analysis to test theoretical predictions in health behavior: An illustration based on meta-analyses of the theory of planned behavior. *Preventive Medicine* 89:154–161.

Hertel, T.W., Golub, A.A., Jones, A.D., O'Hare, M., Plevin, R.J., and Kammen, D.M. 2010. Effects of US Maize Ethanol on Global Land Use and Greenhouse Gas Emissions: Estimating Market-mediated Responses. *BioScience,* 60:223–231.

Hoerl, A.E., and Kennard, R.W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12:55–67.

Hong, J., Shen, Q., and Xue, F. 2016. A multi-regional structural path analysis of the energy supply chain in China's construction industry. *Energy Policy,* 92:56–68.

Iqbal, S., Mahmood, T., Tahira, M.A., Anwar, M., Sarwar, and M. 2003. Path coefficient analysis in different genotypes of soybean (*Glycine max* (L) Merril). *PakistanJournal of Biological Science,* 6:1085–1087.

Jadhav, N.H., Kashid, D.N., and Kulkarni, S.R. 2014. Subset selection in multiple linear regression in the presence of outlier and multicollinearity. *Statistical Methodology,* 19:44–59.

Jones, D.F. 1918. The Effect of Inbreeding and Crossbreeding upon Development. *Proceedings of the National Academy of Sciences of the United States of America* 4:246–250.

Khameneh, M.M., Bahraminejad, S., Sadeghi, F., Honarmand, S.J., and Maniee, M. 2012. Path analysis and multivariate factorial analyses for determining interrelationships between grain yield and related characters in maize hybrids. African Journal of *Agricultural Research,* 7:6437–6446.

Khan, N., and Naqvi, F.N. 2012. Correlation and path coefficient analysis in wheat genotypes under irrigated and non-irrigated conditions. *Asian Journal of Agricultural Sciences,* 4:346–351.

Kumar, K.V., Sudarshan, M.R., Dangi, K.S., and Reddy, S.M. 2013. Character association and path coefficient analysis for seed yield in quality protein maize Zea mays L. *Journal of Research ANGRAU,* 41:153–157.

Kumar, S.V.V., and Babu, D.R., 2015. Character association and path analysis of grain yield and yield components in Maize (*Zea Mays* L.). *Electronic Journal of Plant Breeding,* 6:550–554.

Kumar, T.S., Reddy, D.M., Reddy, K.H., and Sudhakar, P. 2011. Targeting of traits through assessment of interrelationship and path analysis between yield and yield components for grain yield improvement in single cross hybrids of maize (Zea mays L.). *International Journal of Applied Biology and Pharmaceutical Technology,* 2:123–129.

Lee, E.A., and Tollenaar, M. 2007. Physiological basis of successful breeding strategies for maize grain yield. *Crop Science* 47:S-202.

Luz, L.N. da, Santos, R.C. dos, Filho, M., and Albuquerque, P. de 2011. Correlations and path analysis of peanut traits associated with the peg. *Crop Breeding and Applied Biotechnology*, 11:88–95.

Ma, Z., Qin, Y., Wang, Y., Zhao, X., Zhang, F., Tang, J., and Fu, Z., 2015. Proteomic analysis of silk viability in maize inbred lines and their corresponding hybrids. PLoS One 10: e0144050.

Mansfield, E.R., and Helms, B.P., 1982. Detecting Multicollinearity. *The American Statistician,* 36:158–160.

Marquardt, D.W. 1970. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics* 12:591–612.

Mitchell, T.J., and Beauchamp, J.J. 1988. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association,* 83:1023–1032.

Mock, J.J., and Pearce, R.B., 1975. An ideotype of maize. *Euphytica* 24:613–623.

Mohammadi, R., Farshadfar, E., and Amri, A. 2016. Path analysis of genotype x environment interactions in rainfed durum wheat. *Plant Production Science,* 19:43–50.

Mohammadi, S.A., Prasanna, B.M., and Singh, N.N. 2003. Sequential Path Model for Determining Interrelationships among Grain Yield and Related Characters in Maize. *Crop Science* 43:1690–1697.

Nardino, M., Souza, V.Q. de, Baretta, D., Konflanz, V.A., Carvalho, I.R., Follmann, D.N., and Caron, B.O., 2016. Association of secondary traits with yield in maize F_1's. *Ciência Rural* 46:776–782.

Nataraj, V., Shahi, J.P., and Agarwal, V., 2014. Correlation and Path Analysis in Certain Inbred Genotypes of Maize (*Zea Mays* L.) at Varanasi. *International Journal of Innovative Research and Development,* 3:14–17.

Nataraj, V., Shahi, J.P., and Vandana, D. 2015. Character association and path analyses in maize (*Zea mays* L.). *Environment and Ecology,* 33:78–81.

Niles, H.E. 1922. Correlation, Causation and Wright's Theory of "Path Coefficients." *Genetics,* 7:258–273.

Nishii, R. 1984. Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression. *Annals of Statistics,* 12:758–765.

Nogueira, A.P.O., Sediyama, T., Sousa, L.B. de, Hamawaki, O.T., Cruz, C.D., Pereira, D.G., and Matsuo, É., 2012. Análise de trilha e correlações entre caracteres em soja cultivada em duas épocas de semeadura. *Bioscience Journal,* 28:877-888.

Norris, D., Brown, D., Moela, A.K., Selolo, T.C., Mabelebele, M., Ngambi, J.W., and Tyasi, T.L. 2015. Path coefficient and path analysis of body weight and biometric traits in indigenous goats. *Indian Journal of Animal Research,* 49:573–578.

O'brien, R.M., 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Qual Quant,* 41:673–690.

Ojo, D.K., Omikunle, O.A., Oduwaye, O.A., Ajala, M.O., Ogunbayo, S.A., 2006. Heritability, character correlation and path coefficient analysis among six inbred-lines of maize (*Zea mays* L.). *World Journal of Agricultural Sciences*, 2:352–358.

Olivoto, T., Souza, V.Q. de, Carvalho, I.R.C., Nardino, M., and Follmann, D.N. 2015. Análise de trilha para caracteres relacionados ao crescimento de mudas de pepineiro. *Enciclopédia Biosfera,* 11:69–80.

Önder, H., and Abaci, S.H. 2015. Path analysis for body measurements on body weight of saanen kids. *Kafkas Üniversitesi Veteriner Fakültesi Dergisi,* 21:351–354.

Pearson, K. 1920. Notes on the history of correlation. *Biometrika,* 13:25–45.

Petraitis, P.S., Dunham, A.E., and Niewiarowski, P.H. 1996. Inferring multiple causality: the limitations of path analysis. *Functional Ecology*10:421–431.

Reddy, V.R., Jabeen, F., Sudarshan, M.R., and Rao, A.S. 2012. Studies on genetic variability, heritability, correlation and path analysis in maize (*Zea mays* L.) Over locations. *International Journal of Applied Biology and Pharmaceutical Technology,* 4:196–199.

Rigon, J.P.G., Capuani, S., Brito Neto, J.F. de, Rosa, G.M. da, Wastowski, A.D., and Rigon, C.A.G. 2012. Dissimilaridade genética e análise de trilha de cultivares de soja avaliada por meio de descritores quantitativos. *Revista Ceres*, 59:233–240.

Sa, K.J., Park, J.Y., Woo, S.Y., Ramekar, R.V., Jang, C.-S., and Lee, J.K. 2014. Mapping of QTL traits in maize using a RIL population derived from a cross of dent maize × waxy maize. *Genes& Genomics,* 37:1–14.

Saleem, U.S., Subhani, G.M., Ahmad, N., Rahim, M., and Ali, M.A. 2007. Correlation and path coefficient analysis in maize (*Zea mays* L.). *Journal of Agriculture Research* 45:177–183.

Shrivastava, M., and Sharma, K. 1976. Analysis of path coefficients in rice. *Zeitschrift fuer Pflanzenzuechtung*, 77:174–177.

Shull, G.H. 1908. The composition of a field of maize. Journal of Heredity 4:296–301.

Souza, V.Q. de, Bellé, R., Ferrari, M., de Pelegrin, A.J., Caron, B.O., Nardino, M., Follmann, D.N., and Carvalho, I.R. 2015. Componentes de rendimento em combinações de fungicidas e inseticidas e análise de trilha em soja. *Global Science and Technology*, 8:167-176.

Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., McMullen, M.D., Holland, J.B., and Buckler, E.S. 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetetics,* 43:159–162.

Toebe, M., and Cargnelutti, A. 2013. Multicollinearity in path analysis of maize (Zea mays L.). *Journal of Cereal Science,* 57:453–462

Torres, F.E., Teodoro, P.E., Ribeiro, L.P., Correa, C.C.G., Hernandes, F.B., Fernandes, R.L., Gomes, A.C., and Lopes, K.V. 2015. Correlations and path analysis on oil content of castor genotypes. *Bioscience Journal,* 31:1363–1369.

Wold, S., Ruhe, A., Wold, H., Dunn, I., W., 1984. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. SIAM J. Sci. and Stat. Comput. 5, 735–743. doi:10.1137/0905052

Wright, S. 1921. Correlation and causation. *Journal of agricultural research,* 20:557–585.

Wright, S. 1923. The Theory of Path Coefficients a Reply to Niles's Criticism. *Genetics,* 8:239–255.

Xu, L., Lin, T., Xu, Y., Xiao, L., Ye, Z., and Cui, S. 2014. Path analysis of factors influencing household solid waste generation: a case study of Xiamen Island, China. *Journal of Material Cycles and Waste Management,* 18:377–384.

****