







Novel approach to the selection of *Psidium guajava* genotypes using latent traits to bypass multicollinearity

Flavia Alves da Silva* , Caio Cezar Guedes Correa , Beatriz Murizini Carvalho , Alexandre Pio Viana , Sandra da Costa Preisigke , Antônio Teixeira do Amaral Júnior 

Universidade Estadual do Norte Fluminense Darcy Ribeiro/
CCTA – Lab. de Melhoramento Genético Vegetal, Av. Alberto
Lamego, 2000 – 28013-602 – Campos dos Goytacazes,
RJ – Brasil.

*Corresponding author <flavia_uems@hotmail.com>

Edited by: Luiz Alexandre Peternelli

Received April 23, 2019

Accepted September 05, 2019

ABSTRACT: Multicollinearity is a very common problem in studies that employ path analysis in agronomic crops, which generates unrealistic results and erroneous interpretations. This study was aimed at assessing the path analysis in data obtained from guava tree full-sib based on modelling multiple regressions applying latent variables to neutralize the effects of multicollinearity. Seven explanatory variables were measured – fruit mass (FM), fruit length (FL), fruit diameter (FD), mesocarp thickness (MT), peel thickness (PT), pulp mass (PM), total number of fruits (NTF) –, plus the main dependent variable, total yield per plant (YIELD). In accordance with the multicollinearity scenario, eleven values were tested with the addition of the constant K to the diagonal of the correlation matrix $X'X$. Path analysis was applied in two models: all the explanatory variables with direct effect on the dependent one and another model with multiple regression with more than one chain and the presence of latent variables. The path analysis in the multivariate methodology of structural equation modelling (SEM), which uses latent variable prediction, provided better results than the traditional and ridge path analyses.

Keywords: SEM methodology, trail crest, correlation

Introduction

One of the biggest challenges in the cultivation of fruit trees is the high level of investment required and the ability to model associations between traits. The correlation between traits can turn the selection of superior materials into a costly and time-consuming activity. It includes certain steps that, apart from requiring good planning, and financial and manpower resources, mainly require time to obtain the genotypes in the reproductive phase (Grattapaglia and Resende, 2011).

One way is the indirect selection of variables using ridge path analysis. A number of research studies were conducted on fruit trees to identify the real relationships of cause and effect that apply to the ridge path analysis (Kherwar and Usha, 2016; Patel et al., 2015). Many effects close to zero can be observed in the results, which does not mean a lack of relationships between the variables. This is mainly due to multicollinearity which is the existence of a strong relationship between the explanatory variables, and makes an interpretation of the results difficult or non-variable (Farrar and Glauber, 1967; Hair et al., 1995). Multicollinearity can be easily detected by observing the eigenvalues of the matrix $(X'X)$. The ratio between the absolute values of the highest and the lowest eigenvalues gives an idea of the collinearity, as well as the diagonal elements of the matrix $(X'X)^{-1}$ (Montgomery et al., 2012).

In addition to these considerations, the implementation of combined techniques, in which the associations require multivariate statistical procedures of structural equation modelling (SEM) by means of clusters using multiple regressions (Mueller and Hancock, 2018), produces more reliable results for biological phenomena, by manipulating missing data (Enders and Mansolf, 2018)

and estimating latent variables (not observed) (Hair et al., 2014). This equation modelling has been successfully applied in plants, mainly in ecology and evolutionary biology studies (Lefcheck, 2016; Pugesek et al., 2003).

Using the latent variables approach allows us to leave the plastered model of the common path, and group variables with similar characteristics. This grouping, formulated using a variable created mathematically in the model (not observed, known as "latent"), depends on knowledge of the studied biological phenomenon in order for it to make sense. As such, the purpose of this study was to apply path analysis in data from guava full-sibs by means of multiple regression modelling using latent variables aimed at neutralizing the effects of multicollinearity.

Material and Methods

Experimental procedures and genetic material

The data applied here were from experiments performed at Campos dos Goytacazes, in the state of Rio de Janeiro State, Brazil (21°08'02" S, 41°40'47" W, altitude of 18 m). Seventeen full-sib families of guava tree were assessed from controlled crosses between parents.

The experiment was conducted using a randomized block design with two replicates and 24 individuals for each family. Cultural traits recommended for guava culture were respected (Quintal et al., 2017).

Data collection

Seven explanatory variables were measured for each individual – fruit mass (FM) and pulp mass (PM) with the help of a semi analytical balance expressed

in gr; fruit length (FL), fruit diameter (FD), mesocarp thickness (MT), peel thickness (PT) and pulp thickness (PT) and the aid of a pachymeter with the data expressed in mm; total number of fruits (NTF) were counted in the harvest period, counting all fruits harvested from each plant (identifying which fruits were viable or not), plus the main variable, total yield per plant (YIELD). This was carried out during the harvest period, when all the fruits harvested in each plant were weighed in semi-analytical bullet and expressed in g. Five observations of all variables were made except for NTF and YIELD, for which just one observation per individual was made.

Statistical analyses

Pearson linear correlation coefficients (phenotypic correlations) were calculated for the eight variables and measured in the two following ways: (i) using only the number of paired observations in all variables whereby the yield and total number of fruits were measured once per plant per harvest while the other variables were measured five times in each plant, which generated numbers for different observations for the variables, which resulted in 408 observations limited by the variables YIELD and NTF, and (ii) applying all available observations specific to each variable ($408 \leq 1.569$). Subsequently, a matrix $X'X$ of n order was generated (in which: n = number of explanatory variables) with the correlation coefficients and another matrix $X'Y$ of $n \times 1$ dimension (correlation coefficients of the explanatory variables with the dependent variable, YIELD).

A multicollinearity diagnosis was made to obtain the diagonal of the matrix $X'X^{-1}$. It was considered that severe multicollinearity had been reached when the values of the variance inflation factor (VIF) were greater than 10 (Hair et al., 1995). Where there was collinearity, a new diagnosis was made, testing 11 values in the addition of a constant K to the diagonal of the correlation matrix $X'X$ ($K = 0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, \text{ and } 0.1$) to try to reduce the variance associated with the least squares estimator in the path analysis and stabilize the coefficients. These values were chosen by using a wide range of values, hoping that one of them could decrease multicollinearity.

Next, the path analyses were plotted for all situations (paired data, all observations, and admitting values for K), using the system of normal equations $X'X\hat{\beta} = X'Y$ to estimate the direct and indirect effects of the explanatory variables on the dependent variable. The following model was applied:

$$YIELD = \hat{\beta}_1 FM + \hat{\beta}_2 FL + \hat{\beta}_3 FD + \hat{\beta}_4 MT + \hat{\beta}_5 PT + \hat{\beta}_6 PM + \hat{\beta}_7 NTF + e \quad (1)$$

in which: $\hat{\beta}_{1...7}$ were the estimators of the direct effects of the variables FM, FL, FD, MT, PT, PM, and NTF, respectively, on the dependent variable YIELD, acting indirectly through their effects on the other explanatory variables, and e the residual term of model. The equation

determination coefficient ($R^2 = 1 - (\sum_{i=1}^n (Y_i - \bar{Y}_i)^2 / e)$) and the effect of the residue variable ($e = \sum_{i=1}^n (Y_i - \bar{Y}_i)^2$) of the path analysis were calculated for all situations. Next, we proceeded to define a latent variable model, whereby a latent variable grouped trait that we believe to have similar characteristics, was coincident with the biological phenomenon. The analysis was conducted again by adjusting the model for the latent variables as follows:

$$YIELD = \hat{\beta}_1 NTF + \hat{\beta}_2 FM \quad (2)$$

$$FM = \hat{\beta}_3 L1 + \hat{\beta}_4 L2 \quad (3)$$

$$L1 = \hat{\beta}_5 PM + \hat{\beta}_6 MT \quad (4)$$

$$L2 = \hat{\beta}_7 FL + \hat{\beta}_8 FD + \hat{\beta}_9 PT \quad (5)$$

in which: $\hat{\beta}_1$ and $\hat{\beta}_2$ were the direct effects of the variables NTF and FM on the main dependent variable YIELD; $\hat{\beta}_3$ and $\hat{\beta}_4$ the direct effects of the latent variables L1 and L2 on FM; $\hat{\beta}_5$ and $\hat{\beta}_6$ the effects of the variables PM and MT on the latent variable L1; $\hat{\beta}_7$, $\hat{\beta}_8$, $\hat{\beta}_9$ and were the direct effects of the variables FL, FD, and PT on the latent variable L2. As regards the modeling latent variables, the same assumptions are required for path analysis, as well as the distribution of errors $NID(0, \sigma^2)$. The illustrative causal diagram of the models can be seen in Figure 1.

All analyses were carried out by means of the R software (R, version 3.5.0), using the following packages: *biotools* 3.1 (Silva et al., 2017), *semPlot* 1.1 (Epskamp, 2015), and *lavaan* 0.6 (Rosseel, 2012).

Results and Discussion

Pearson linear correlation was estimated for the eight variables, applying only the paired data and all available observations for the variables (Table 1). Afterwards, the correlation between those two matrices, using pairwise data and a different number of observations ($r = 0.69^{**}$) was obtained and the Mantel test (0.27434^{+*}) was conducted. There was no statistical differentiation at 1 % level of probability for the correlation estimates either for the t test or the critical level of Mantel. Furthermore, the use of different observations resulted in a significant difference (30%), which does not produce a true biological differential effect, considering that the progenies are descended from the same ancestral population.

For most estimates, the magnitudes and senses of the correlations were maintained. Nevertheless, it was noted that, in these matrices, a number of correlations were altered, such as that between the variables FL and FD, in which it was possible to identify an increase in correlation ($r = -0.0062$ for 0.6581) when applying more observations. Other examples were found for the variables FM and FD ($r = 0.0253$; 0.9012); FM and PM (0.2371 ; 0.9534); FD and FL (-0.0062 ; 0.6581), in which

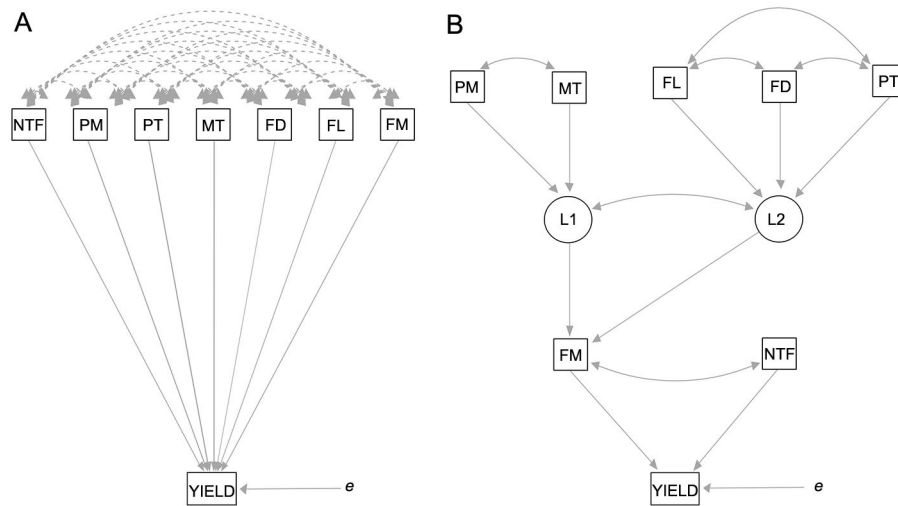


Figure 1 – (A) – Causal diagram with the direct effects (unidirectional continuous lines) of the following explanatory variables Fruit Mass (FM), Fruit Length (FL), Fruit Diameter (FD), Mesocarp Thickness (MT), Peel Thickness (PT), Pulp Mass (PM), Number of Total Fruit (NTF) on the dependent variable Total Yield per Plant (YIELD) and indirect effects (bidirectional dashed lines) of the explanatory variables on the dependent variable for the adjustment of the ordinary path model. (B) – Causal diagram showing the direct (unidirectional arrows) and indirect effects (bidirectional arrows) among the observed variables (squares) of the model using the adjustment of models with latent variables expressed by circles: latent variable 1 (L1) and latent variable 2 (L2). This is the effect of the residue variable on the dependent variable YIELD.

Table 1 – Coefficients of Pearson linear correlation among eight variables of guava tree. In the upper diagonal, correlations obtained from all available observations for each variable. In the lower diagonal, correlations obtained from 408 pairs of observations. Fruit Mass (FM – 1.569 observations), Fruit Length (FL – 1.569 observations), Fruit Diameter (FD – 1.569 observations), Mesocarp Thickness (MT – 1.569 observations), Pulp Thickness (PT – 1.569 observations), Pulp Mass (PM – 1.569 observations), Total Yield per Plant (YIELD – 408 observations), and Number of Total Fruits (NTF – 408 observations).

	FM	FL	FD	MT	PT	PM	YIELD	NTF
FM	–	0.7847	0.9012	0.5803	0.1362	0.9534	-0.2839	-0.1934
FL	0.7690	–	0.6581	0.4119	0.1090	0.7704	-0.2362	-0.1697
FD	0.0253	-0.0062	–	0.5938	0.0525	0.8878	-0.3471	-0.2142
MT	0.6184	0.4388	0.0187	–	0.0833	0.6630	-0.2117	-0.1076
PT	0.1430	0.1336	-0.0112	0.0866	–	0.1383	0.1700	0.0579
PM	0.2371	0.2025	0.0039	0.1940	0.0224	–	-0.2788	-0.1812
YIELD	-0.2839	-0.2362	-0.3471	-0.2117	0.1700	-0.2788	–	0.4861
NTF	-0.1934	-0.1697	-0.2142	-0.1076	0.0579	-0.1812	0.5231	–

high positive correlations were expected, but they did not materialize when reduced quantum of observations were used.

The analysis continued with the multicollinearity report on the basis of the variance inflation factor (VIF) from the diagonal of the correlation matrix $X'X^{-1}$ using the complete data, in which collinearity was considered for the variables that showed values higher than 10. In the results, collinearity problems in the variables can be seen, in which the variables FM and PM showed VIF higher than the limit (16.47 and 15.30, respectively). The multicollinearity was confirmed, and a constant was added to the diagonal of the matrix $X'X$ to obtain the lowest possible value of that constant, which stabilizes the path coefficients.

Because of the effects of the constant value on the variables (Figure 2), with the increase in the constant K , the residue effect also increased. This effect is inversely proportional to the value of the regression equation determination coefficient (R^2), as with the increase in the values given to K , the values of R^2 decreased (Figure 2). The first value of R^2 was 0.0, which takes into account the path analysis without the addition of the constant. The initial value for R^2 in that scenario was 0.35, and the residue effect, of 0.802. The lowest value for the constant K that stabilized the variances ($VIF < 10$) was 0.03, in which the variables that displayed problems of variance inflation increased to the values of 9.56 and 9.11 for FM and PM, respectively, resolving the multicollinearity problem. However, the value of the

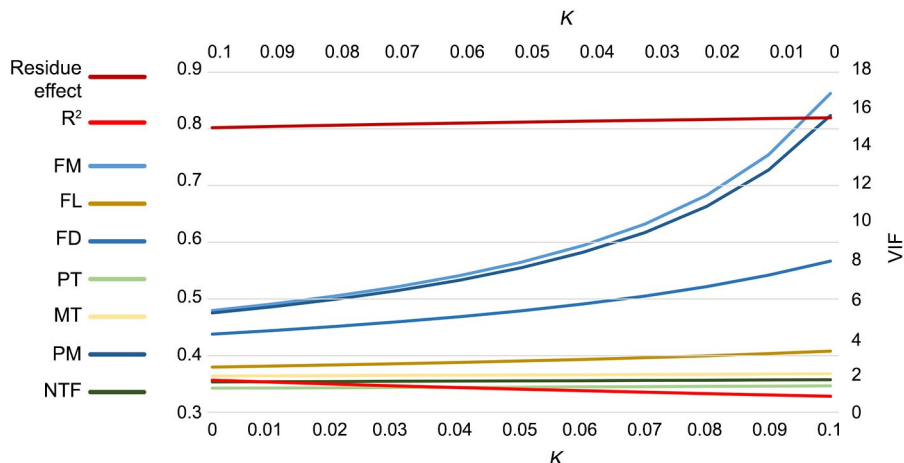


Figure 2 – Values of the regression equation determination coefficient (R^2) and values for the variable residue effect on YIELD (they follow the scale noted to the left – $0.3 > 0.9$ – and the lower x-axis). The other lines are the variance inflation values for the following variables: Fruit Mass (FM), Fruit Length (FL), Fruit Diameter (FD), Mesocarp Thickness (MT), Peel Thickness (PT), Pulp Mass (PM), Number of Total Fruits (NTF), depending on the values of constant K (following the scale to the right – $0 > 18$ – and the upper x-axis).

equation determination coefficient decreased ($R^2 = 0.34$), and consequently, the model now explains less of the data. An increase in the residue effect on the dependent variable (0.808) was also observed.

In this study, in which the implementation of a value to correct the matrix $X'X$ generated cause and effect relationships much closer to zero corroborates previously published results such as those found in studies of multicollinearity in maize (Olivoto et al., 2017). This has also been seen in studies where many variables have been applied which study traits of commercial interest in guava trees (Kherwar and Usha, 2017).

On the basis of the ridge path analysis, using the value of 0.03 for the constant K (Table 2), it was noted that, generally, there were values close to zero both for the direct and the indirect effects; this can be seen in the estimates of the indirect influences of the variables FL, MT, PT, and PM, on the variable NTF, together with its effects on the YIELD, with values of 0.009; 0.006; 0.009; and -0.009, respectively. The greatest direct effect was for the variable NTF on YIELD (0.409), followed by the most pronounced direct influences of PT and FD on the YIELD, with corresponding values of 0.152 and -0.290.

It is worth noting that, between the cause and effect relations, the most significant estimates were given by the variable FD for the variables FM and PM (-0.261 and -0.257). With this result, the association of these estimates with biological effects becomes unfeasible, as the increase in fruit diameter – measured by a longitudinal cut in the fruit – was able to generate fruits with smaller mass and smaller pulp mass. However, this would not be reliable, since all fruits have a spherical or pear shape; thus, larger diameters necessarily imply larger mass.

As for trees, the great majority had values close to zero for both the direct and indirect effects in the case of variables associated with the plant growth, and

Table 2 – Ridge path analysis ($K = 0.03$) for the following variables: Fruit Mass (FM), Fruit Length (FL), Fruit Diameter (FD), Mesocarp Thickness (MT), Peel Thickness (PT), Pulp Mass (PM), Number of Total Fruit (NTF) seen in guava trees with their indirect effects on the dependent variable Total Yield per Plant (YIELD). Direct effects are shown in the main diagonal.

*	FM	FL	FD	MT	PT	PM	NTF
FM	0.058	-0.043	-0.261	-0.030	0.021	0.049	-0.079
FL	0.046	-0.055	-0.191	-0.021	0.017	0.039	-0.069
FD	0.052	-0.036	-0.290	-0.031	0.008	0.045	-0.088
MT	0.034	-0.023	-0.172	-0.052	0.013	0.034	-0.044
PT	0.008	-0.006	-0.015	-0.004	0.152	0.007	0.024
PM	0.055	-0.042	-0.257	-0.034	0.021	0.051	-0.074
NTF	-0.011	0.009	0.062	0.006	0.009	-0.009	0.409

*The effect of the residue variable $e = 0.81$; the model determination coefficient $R^2 = 0.34$.

for variables assessed in the flowers and yield variables (Patel et al., 2015). Clearly it has had inappropriate results with the biological phenomena, confirming the need to improve the technique.

Another answer provided regarding the limitation of ordinary path analysis is that, when no data treatment is undertaken (correction factor in the matrix $X'X$, data transformation, standardization, and so forth) and data has been given with collinear variables, the incidence of coefficients that exceed the expected limit is common ($-1 < 1$), such as was seen in the study results of Santos et al. (2017), who researched the cause and effect relation between variables of plant growth and yield.

By applying the path analysis methodology using latent variables, multiple regression models were implemented, arranging the path in more than one chain (Figure 3). The latent variables were set in a chain level and suffered the influence of the variables during

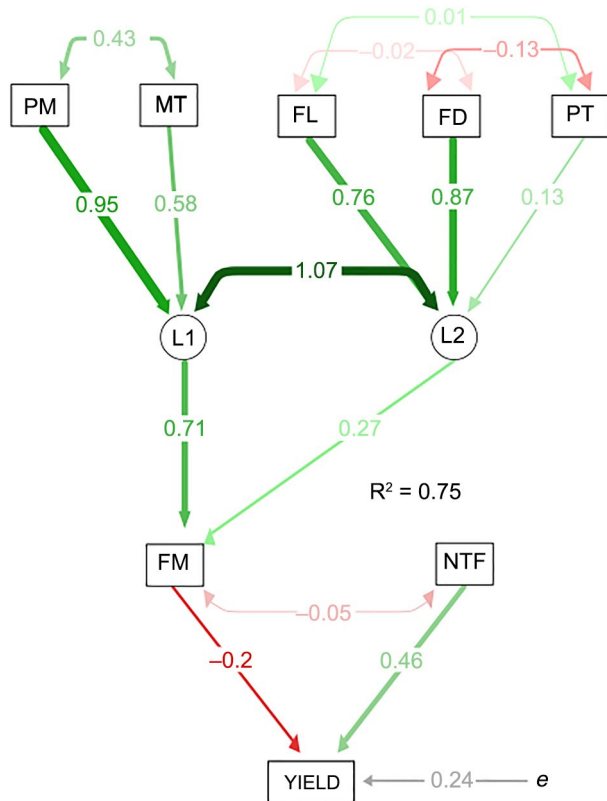


Figure 3 – Causal diagram of the path analysis with latent variables in nine explanatory variables (FM = Fruit Mass, FL = Fruit Length, FD = Fruit Diameter, MT = Mesocarp Thickness, PT = Peel Thickness, PM = Pulp Mass, NTF = Number of Total Fruit, being two latent ones, L1 = latent variable 1 and L2 = latent variable 2), and the main variable Total Yield per Plant (YIELD), in which e is the effect of the residue variable on the dependent variable YIELD, and R² the model determination coefficient. Bidirectional arrows show correlation between the variables, and unidirectional arrows indicate a direct effect on the direction of the arrow; the green color evidences positive correlation and the red negative correlation.

the assessments of the experiments and with a better biological reasoning, rather than observing the effects of all variables correlated with each other and with the dependent one.

The variables PM, PT, FL, FD, and PT – obtained by assessing the fruits – converged their effects in the path on the variable FM. An expressive gain was achieved in the explanatory power of the model when observing that the determination coefficient went from R² = 0.3464 to 0.75, for the ridge path analysis (K=0.03) and multiple regression models with latent variables, respectively. Improvements can also be noticed in the residue effect on the variable YIELD, which had a reduced magnitude down from 0.8084 to 0.24. Expressive improvements in the estimates were also achieved by Dehghani et al. (2009) in the implementation of multiple equations for the path analysis in traits of economic interest in melon.

With great similarity, these authors reported the same problems found in this study regarding the guava tree; they obtained satisfactory results after the appropriate arrangement tests of the effects on the variables.

In the models developed herein, the variables PM and MT exert a direct effect on the latent variable L1, in which there is a strong influence of the variable PM (0.95) together with the effect of the variable MT (0.58), both with strong positive effects that, combined, produce an influence (0.71) greater than the latent variable L2 on the FM. These results confirm what was expected for relationships in which fruits with greater mass and pulp thickness (mesocarp) clearly need to be larger, resulting in a greater fruit mass.

The value of 1.07 between latent traits L1 and L2 indicates multicollinearity because the value exceeds the unit (parametric space for correlation and path analysis). This had already been expected because there are many traits that control the two latent variables, and the theoretical relationship between the two would be very strong, which would also serve as a buffer effect in the model. However, since these variables are not studied in their relationships, and only serve to connect the model, there is no problem with the multicollinearity between them.

The variables FL, FD, and PT influence the latent variable L2. The greatest effect is seen in the variable FD (0.87) followed by the variable FL (0.76) and the low influence of the variable peel thickness (0.13). All these variables together result in the effect that L2 expresses on the fruit mass (0.27). This effect on the mass fruit is smaller than the one observed in the variable L1 (0.71).

No less important is the fact that these variables can be chosen to modify the fruit shape (varying between spherical and pear). The variable PT, despite the little influence, may be of interest for yields related to fruit shelf-life, in which a thicker peel can extend the fruit shelf-life because of its greater resistance to the infusion of O₂ into the fruit, which would increase the deterioration rate (Teixeira et al., 2016). Negative correlation between PT and FD (0.13) was observed, which, despite being low, is perfectly acceptable from a biological point of view; it is still a result that should be closely assessed in case table fruits are desired, considering that the selection of genotypes that yield great fruit can have a thinner peel.

In general, all those variables of the third path chain can be indirectly controlled by cultural traits. In addition to providing good local control both appropriate pruning and maintaining the ideal number of branches are required to influence the number of fruits, since, in each crop, a branch that has a new bud results in up to three fruits. If an excessive number of reproductive buds is maintained the plant will need to distribute the photo-assimilated ones among more fruits which would result in smaller fruits (Serrano et al., 2008).

This experience describes the negative indirect effect of the fruit mass on the number of fruits (0.05),

which, despite being small, when considering a mean yield between 40 and 65 t ha⁻¹, in the end, significant differences can be calculated. The negative direct effect of the fruit mass on the yield is also related to this event, in which a plant that yields a few fruits produces large fruits with a larger mass; nevertheless, a plant that yields more fruits also produces smaller fruits, but the sum of the mass is greater, and thus the yield is higher.

Conclusions

The path analysis with the implementation of the SEM methodology, which uses latent variable prediction, succeeded in delivering better results than ordinary path analysis and ridge path analysis. It is possible to indirectly choose the variable fruit mass by means of the pulp mass and fruit diameter of the variables. For indirect selection of the variable yield, the genotypes should be selected according to the number of fruits per variable.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) also assisted the execution of this experiment.

Authors' Contributions

Conceptualization: Silva, F.A.; Viana, A.P. **Data acquisition:** Silva, F.A.; Carvalho, B.M.; Preisigke, S.C. **Data analysis:** Correa, C.C.G. **Design of methodology:** Silva, F.A.; Amaral Júnior, A.T. **Software development:** Silva, F.A.; Correa, C.C.G. **Writing and editing:** Silva, F.A.; Correa, C.C.G.; Carvalho, B.M.

References

- Dehghani, H.; Feyzian, E.; Rezai, A.; Jalali, M. 2009. Correlation and sequential path model for some yield-related traits in melon (*Cucumis melo* L.). *Journal of Agricultural Science and Technology* 11: 341-353.
- Enders, C.K., Mansolf, M. 2018. Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods* 23: 76.
- Epskamp, S. 2015. *semPlot*: unified visualizations of structural equation models. *Structural Equation Modeling* 22: 474-483. <https://doi.org/10.1080/10705511.2014.937847>
- Farrar, D.E.; Glauber, R.R. 1967. Multicollinearity in regression analysis: the problem revisited. *The Review of Economics and Statistics* 49: 92-107.
- Grattapaglia, D.; Resende, M.D.V. 2011. Genomic selection in forest tree breeding. *Tree Genetics Genomes* 7: 241-255.
- Hair, J.F.; Anderson, R.E.; Tatham, R.L.; Black, W.C. 1995. *Multivariate data analyses with readings*. 4ed. Pearson Education, Hoboken, NJ, USA.
- Hair, J.F.; Sarstedt, M.; Hopkins, L.; Kuppelwieser, V.G. 2014. Partial least squares structural equation modeling (PLS-SEM): an emerging tool in business research. *European Business Review* 26: 106-121.
- Kherwar, D.; Usha, K. 2016. Genetic variations, character association and path analysis studies in guava (*Psidium guajava* L.) for bioactive and antioxidant attributes. *Indian Journal of Plant Physiology* 21: 355-361.
- Kherwar, D.; Usha, K. 2017. Character association and path analysis studies in guava (*Psidium guajava* L.) for bioactive and antioxidant attributes. *Progressive Horticulture* 49: 30-35. <https://doi.org/10.5958/2249-5258.2017.00007.0>
- Lefcheck, J.S. 2016. *piecewiseSEM*: piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution* 7: 573-579.
- Montgomery, D.C.; Peck, E.A.; Vining, G.G. 2012. *Introduction to Linear Regression Analysis*. John Wiley, Hoboken, NJ, USA.
- Mueller, R.O.; Hancock, G.R. 2018. Structural equation modeling. p. 445-456. In: Hancock, G.R.; Mueller, R.O.; Stapleton, L.M., eds. *The reviewer's guide to quantitative methods in the social sciences*. Routledge, New York, NY, USA.
- Olivoto, T.; Souza, V.Q.; Nardino, M.; Carvalho, I.R.; Ferrari, M.; Pelegrin, A.J.; Szareski, V.J.; Schmidt, D. 2017. Multicollinearity in path analysis: a simple method to reduce its effects. *Agronomy Journal* 109: 131-142.
- Patel, R.; Maiti, C.; Deka, B.; Vermav, V.; Deshmukh, N.; Verma, M. 2015. Genetic variability, character association and path coefficient study in guava (*Psidium guajava* L.) for plant growth, floral and yield attributes. *International Journal of Bio-Resource and Stress Management* 6: 457-466. <https://doi.org/10.5958/0976-4038.2015.00068.8>
- Pugesek, B.H.; Tomer, A.; Von Eye, A. 2003. *Structural Equation Modeling: Applications in Ecological and Evolutionary Biology*. Cambridge University Press, Cambridge, UK.
- Quintal, S.S.R.; Viana, A.P.; Campos, B.; Vivas, M.; Amaral Júnior, A.T. 2017. Selection via mixed models in segregating guava families based on yield and quality traits. *Revista Brasileira de Fruticultura* 39: e-866. <https://doi.org/10.1590/0100-29452017866>
- Rosseel, Y. 2012. *lavaan*: an R package for structural equation modeling. *Journal of Statistical Software* 48: 1-36. <https://doi.org/10.18637/jss.v048.i02>
- Santos, P.R.; Costa, P., S.; Viana, A.P.; Cavalcante, N.R.; Sousa, C.M.B.; Amaral Júnior, A.T. 2017. Associations between vegetative and production traits in guava tree full-sib progenies. *Pesquisa Agropecuária Brasileira* 52: 303-310. <https://doi.org/10.1590/S0100-204X2017000500003>
- Serrano, L.A.L.; Martins, M.V.V.; Melo, I.L.; Marinho, C.S.; Tardin, F.D. 2008. Effect of pruning time and intensity on 'Paluma' guava trees, in Pinheiros, ES, Brazil. *Revista Brasileira de Fruticultura* 30: 994-1000. <https://doi.org/10.1590/S0100-29452008000400026>
- Silva, A.R.; Malafaia, G.; Menezes, I.P.P. 2017. Biotools: an R function to predict spatial gene diversity via an individual-based approach. *Genetics and Molecular Research* 16: gmr16029655.
- Teixeira, G.H.; Cunha Júnior, L.C.; Ferraudo, A.S.; Durigan, J.F. 2016. Quality of guava (*Psidium guajava* L. cv. Pedro Sato) fruit stored in low-O₂ controlled atmospheres is negatively affected by increasing levels of CO₂. *Postharvest Biology and Technology* 111: 62-68.