Scientia Agricola
http://dx.doi.org/10.1590/0103-9016-2014-0186

Point of View

Sequential path analysis: what does "sequential" mean?

Marcin Kozak[1]*, Ricardo Antunes Azevedo[2]

[1]Warsaw University of Life Sciences – Dept. of Botany, Nowoursynowska 159 – 02-766 – Warsaw – Poland.
[2]University of São Paulo/ESALQ – Dept. of Genects, Av. Pádua Dias, 11 – 13418-900 – Piracicaba, SP – Brazil.
*Corresponding author <nyggus@gmail.com>

Edited by: Paulo Cesar Sentelhas

Received May 22, 2014
Accepted June 03, 2014

ABSTRACT: Studying relationships among plant and crop traits is crucial for crop scientists to understand complex biological systems that occur in plants and the field. Such knowledge constitutes the basis for more practical information on how to manage breeding and production to provide better or more suitable cultivars, higher yields, lower yield gaps, and resistance to pests etc. To acquire such knowledge, however, representative models of associations between plant and crop traits must be constructed. In path analysis – one of the major methods for analyzing multivariate relationships between quantitative traits – it is important to decide on an appropriate model for these associations, a model that is representative of the corresponding biological phenomena that are of interest to crop researchers. Adopting this "point of view", we asked various questions relating to such model building: (i) how should sequentiality in sequential path analysis be understood? (ii) how should it be interpreted? (iii) how should such sequential models be formulated? We discussed these issues in the context of crop science. Differences in simple and complex (sequential) models of path analysis are presented. Based on crop science examples, we show how important it is to correctly represent the biological relationships for a path analysis model.

Keywords: interpretation, relationships, models, statistics

Analyzing associations between traits lies at the heart of understanding biological phenomena in plant development. A multiplicity of simple and complex, statistical and exploratory methods is used for such analyses (e.g., Kaya et al., 2006; Jaradat, 2007, 2011; Kozak, 2010a, b; Annicchiarico et al., 2013; Valério et al., 2013; Prohens et al., 2013; Moreno et al., 2014; Silva et al., 2014). Some of them, simple Pearson correlations being an example, merely look at how two traits change together. Others, for example, principal component analysis or factor analysis, attempt to simplify the picture of multivariate relationships to the extent that this picture is understandable and interpretable. However, there are methods in use that provide a picture of causal associations between traits, and that are found in both simple and complex systems. Structural equation modeling is one example.

Path analysis – which, these days, is an integral part of structural equation modeling – has proven to be popular in recent decades for the study of association in quantitative traits. Even though immediately after its development by Wright (1921), path analysis was quite severely criticized (Niles, 1922), it still found its way into many scientific disciplines, including crop science (e.g., Kozak and Kang, 2006; Lorencetti et al., 2006; Kozak et al., 2008; Singh et al., 2008; Zeng and Meredith, 2009; Chitra and Rajamani, 2010; Pedersen et al., 2010; Badu-Apraku et al., 2012). A starting point of this popularity in crop science seems to be Dewey and Lu's (1959) paper. Among path analysis models, we can distinguish two types: (i) *simple models,* in which all traits except for the dependent one are set up at the same ontological level, which makes for their being treated as co-related; and (ii) *complex models,* in which traits are set up at different levels, and relations (that is, co-relations or cause-and-effect relationships) between them are to reflect possible biological relations.

Dewey and Lu's (1959) analysis was based upon simple path analysis models, while for some time, the most popular appear to be the complex path analysis models. A simple reason for this change is that complex models can be representative of biological processes in complex biological systems (see, for instance, the so-called "systems biology"), something which cannot be said for simple models. Because of the way they are formulated, complex models are also referred to as sequential models – they reflect sequential development of crop traits and, thus, sequentiality of cause-and-effect associations among them.

Sequential models of path analysis aim to help researchers understand relations among traits that develop in subsequent stages during plant ontogenesis. As with any statistical model, such sequential models should be representative of biological processes that they are built to analyze. Otherwise, they will be just models without biological meaning, which hinder proper interpretation of the biological processes to be studied.

Kozak and Azevedo (2011) discussed how sequential models should be interpreted, and how sequential models should be built – not in terms of estimation (which is a strictly statistical issue), but in terms of model formulation. The way such models are interpreted reflects the way they are built, and the way they are built must reflect the biological processes they are to represent. Thus, model formulation obviously affects whether an interpretation is biologically correct or not.

If one aims to form a sequential path analysis model, one assumes that some traits develop earlier in ontogeny and affect other traits that develop later. In this way a clear cause-and-effect relationship is formed, in which the trait that develops earlier *affects* the trait that develops later or, in more general terms, one trait *affects* another trait.

We know which trait affects which, or rather – we know which trait can affect which, and we aim to study whether this influence is significant or not. This knowledge, based upon biological knowledge of the processes studied, is used to form a basic model that is then used to build a final model (based on statistical criteria, which is not a topic in this paper). Thus, in standard applications in crop science we do not test which direction of a cause-and-effect relationship is true; we assume it to be so and test whether our data provide sufficient proof of this relationship or not. Of course, we do need to have such knowledge; otherwise we could not formulate models for testing. Refer to Kozak and Azevedo (2011) and citations therein for a discussion about this topic.

Unfortunately, as we discussed previously (Kozak and Azevedo, 2011), since 2004 sequential path analyses have suffered from a methodological flaw that affects interpretation of path models formed. It can be found in a dozen or so publications where causal studies based on path analyses have been carried out (a list of such papers can be found in Kozak and Azevedo, 2011). This flawed methodology is still in use. Herein we aim to direct the readers to this particular aspect of path analysis, especially because application of path analysis, and its sequential version in particular, seems to be very important in various applications in crop science.

In the flawed methodology of sequential path analysis, sequentiality is not related to the direction of causal association among traits (as it should be). Although traits are set up in some sequential order in the model, the methodology of this setting is based upon stepwise variable selection in regression analysis (hereinafter, termed stepwise regression) and analysis of the total contribution of the traits to the variation of the dependent variable (Mohammadi et al., 2003). In short, the stepwise regression selects from the whole set of independent variables a set of those predictors which influence the dependent variable the most. This approach is based upon a classical multiple linear regression model, which means that it ignores the structure of cause-and-effect relationships among the predictors.

A proper approach to analyzing cause-and-effect relationships among crop traits (bot only in path analyses, but also in other methods) assumes that sequentiality in a fitted model reflects sequentiality in the biological sense. This means that the height of a cereal plant can affect grain yield (in a given season), but the opposite relationship is very unlikely. Thus, before estimating a model of cause-and-effect relationships among crop traits, possible paths are set up based on the knowledge of the biology of the processes being studied. These two concepts of sequentiality are quite different and it is the biological concept of sequentiality which is normally intended when performing path analysis in crop research. As follows from our discussion in Kozak and Azevedo (2011), results of both these approaches can often be similar, but this is not always the case and usually it is simply luck if proper conclusions are reached based on improper methodology. Without this luck, such results can lack biological sense, and treating regression-like sequentiality as biological sequentiality can lead to strange and incorrect results.

Is the following question biologically valid? In maize (*Zea mays*), can ear height be a determinant of plant height? Quite likely everyone would agree that plant height and ear height are related maize traits, perhaps with high positive correlation. Both are also determinants of grain yield. Nonetheless, the fact that two events are correlated does not mean that one determines the development of the other, or *vice versa*. It is essential to be cautious when interpreting correlated events or traits. Many researchers show that plant height and ear height are controlled by several common QTLs, but also QTLs with individual control, which means that these traits are correlated but one does not "control" the other (Veldboom et al., 1994; Sibov et al., 2003). Moreover, in terms of their development, too, one could not choose ear height as a determinant of plant height – the process of shaping ears does, in fact, finalize before finalizing plant height. So, ear height should *not* be considered a determinant of plant height.

In literature, however, we might find a conclusion that ear height affects maize plant height, with high significant correlation ($r > 0.7$), in both drought and low-nitrogen environments (Badu-Apraku et al., 2012). Careful examination of the methodology used to draw the above conclusion gives rise to a conclusion that it was due to the way the sequential path model was built. The corresponding model was derived from a method based on model fitting in which a final model is selected based on multiple regression analysis along with the analysis of the variables' total contribution to the variation of the dependent variable. Unfortunately, in this particular situation, the fitted model does not reflect phenomena that we could consider biologically reasonable.

We provided more similar examples on the use of path analysis in Kozak and Azevedo (2011) and discuss more about what is important when studying complex models. In particular, we asked whether maize thousand kernel weight can affect the number of kernels per row; and whether the number of grains per panicle (directly) and thousand-grain weight can affect rice plant height (a similar question to that asked above for maize). It is worth emphasizing, however, that some of the authors were lucky to provide biologically valid models based on improper methodology.

Our aim from this point of view was to emphasize that path analysis should always be carefully con-

ducted and only biologically reasonable models should be considered. Of course, the very same conclusion can be drawn in relation to *every single statistical method* applied in crop research. The point, however, is that *any* statistical analysis should reflect the biological phenomena studied and thus statistics should be treated as a tool supporting interpretation. Sophisticated statistics will be useless if only used for themselves. Sometimes a simple and straightforward analysis (non-statistical methods included, such as those based on visualization techniques, e.g., Kozak, 2010b; Wnuk et al., 2013) is much better than a complex one, particularly in situations when the former does what it is expected to, while the latter mainly shows the skills of the analyst. It is not to say, of course, that one should avoid employing complex statistical methods; sometimes one should. It is to say, however, that employing them makes sense when simpler methods are insufficient to cope with the data and/or research questions. We must never forget why statistics is used in crop science, namely, to help understand biological phenomena.

References

Annicchiarico, P.; Pecetti, L.; Tava, A. 2013. Physiological and morphological traits associated with adaptation of lucerne (*Medicago sativa*) to severely drought-stressed and to irrigated environments. Annals of Applied Biology 162: 27-40.

Badu-Apraku, B.; Akinwale, R.O.; Franco, J.; Oyekunle, M. 2012. Assessment of reliability of secondary traits in selecting for improved grain yield in drought and low-nitrogen environments. Crop Science 52: 2050-2062.

Chitra, R.; Rajamani, K. 2010. Character association and path analysis in glory lily (*Gloriosa superba* L.). Communications in Biometry and Crop Science 5: 78-82.

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

Jaradat, A.A. 2007. Predictive grain yield models based on canopy structure and structural plasticity. Communications in Biometry and Crop Science 2: 74-89.

Jaradat, A.A. 2011. Polymorphism, population structure, and multivariate relationships among secondary traits in open-pollinated corn heterotic groups. Communications in Biometry and Crop Science 6: 4-20.

Kaya, Y.; Akcura, M.; Ayranci, R.; Taner, S. 2006. Pattern analysis of multi-environment trials in bread wheat. Communications in Biometry and Crop Science 1: 63-71.

Kozak, M. 2010a. Basic principles of graphing data. Scientia Agricola 67: 483-494.

Kozak, M.; Azevedo, R.A. 2011. Does using stepwise variable selection to build sequential path analysis models make sense? Physiologia Plantarum 141: 197-200.

Kozak, M.; Kang, M.S. 2006. Note on modern path analysis in application to crop science. Communications in Biometry and Crop Science 1: 32-34.

Kozak, M.; Bocianowski, J.; Rybinski, W. 2008. Selection of promising genotypes based on path and cluster analyses. The Journal of Agricultural Science 146: 85.

Kozak, M. 2010b. Use of parallel coordinate plots in multi-response selection of interesting genotypes. Communications in Biometry and Crop Science 5: 83-95.

Lorencetti, C.; Carvalho, F.I.F.D.; Oliveira, A.C.D.; Valério, I.P.; Hartwig, I.; Benin, G.; Schmidt, D.A.M. 2006. Applicability of phenotypic and canonic correlations and path coefficients in the selection of oat genotypes. Scientia Agricola 63: 11-19.

Mohammadi, S.A.; Prasanna, B.M.; Singh, N.N. 2003. Sequential path model for determining interrelationships among grain yield and related characters in maize. Crop Science 43: 1690-1697.

Moreno, C.; Mancebo, I.; Tarquis, A.M.; Moreno, M.M. 2014. Univariate and multivariate analysis on processing tomato quality under different mulches. Scientia Agricola 71: 114-119.

Niles, H.E. 1922. Correlation, causation and Wright's theory of "path coefficients". Genetics, 7: 258-273.

Pedersen, P.; Tylka, G.L.; Mallarino, A.; Macguidwin, A.E.; Koval, N.C.; Grau, C.R. 2010. Correlation between soil plant height, population densities, and soybean yield. Crop Science 50: 1458-1464.

Prohens, J.; Whitaker, B.D.; Plazas, M.; Vilanova, S.; Hurtado, M.; Blasco, M.; Gramazio, P.; Stommel, J.R. 2013. Genetic diversity in morphological characters and phenolic acids content resulting from an interspecific cross between eggplant, *Solanum melongena*, and its wild ancestor (*S. incanum*). Annals of Applied Biology 162: 242-257.

Silva, A.R.D.; Cecon, P.R.; Dias, C.T.D.S.; Puiatti, M.; Finger, F.L.; Carneiro, A.P.S. 2014. Morphological phenotypic dispersion of garlic cultivars by cluster analysis and multidimensional scaling. Scientia Agricola 71: 38-43.

Singh, A.K.; Singh, H.P.; Singh, S.P.; Kalra, A. 2008. Genetic variability and correlation studies for selection criteria in Safed Musli (*Chlorophytum borivilianum*, Santapau). Communications in Biometry and Crop Science 3: 67-71.

Wright, S. 1921. Correlation and causation. Journal of Agricultural Research 20: 557-585.

Wnuk, A.; Górny, A.G.; Bocianowski, J.; Kozak, M. 2013. Visualizing harvest index in crops. Communications in Biometry and Crop Science 8: 48-59.

Valério, I.P.; Carvalho, F.I.F.D.; Benin, G.; Silveira, G.D.; Silva, J.A.G.D.; Nornberg, R.; Hagemann, T.; De Souza Luche, H.; Oliveira, A.C.D. 2013. Seeding density in wheat: the more, the merrier? Scientia Agricola 70: 176-184.

Zeng, L.; Meredith, W.R. 2009. Associations among lint yield, yield components, and fiber properties in an introgressed population of cotton. Crop Science 49: 1647-1654.