Note on Yield Component Analysis

Marcin Kozak, Wiesław Mądry
Department of Biometry,
Warsaw Agricultural University, Nowoursynowska 159, 02-776 Warsaw, Poland,
e-mail: marcin.kozak@omega.sggw.waw.pl

Summary
A discussion on yield component analysis is presented in the paper. The discussion regards such issues as inappropriateness of a linear model in application to yield component analysis, correct measurement of yield and its components, sequential and nonsequential development of the components, and a choice of a statistical analysis method to provide a view on a relationship between yield and its components.

Key words: determination coefficient, nonsequential development, sequential development, yield, yield components.

1. Introduction

Yield component analysis is a methodology that involves statistical evaluation and interpretation of complex relationships between final crop yield per plant or unit area (plot) and so-called yield components. Yield components are such yield-contributing characters of canopy or plant which product or sum, per respective experimental unit, is equal to the yield value per this unit (e.g., Piepho, 1995; Gołaszewski, 1996; Kozak, 2004). Such yield-contributing characters are called yield components, multiplicative or additive ones, respectively. Consider the simplest example: yield of any crop plant per unit area can be evaluated by multiplying its two components: number of plants on this unit area and average yield of one plant from this unit area. Hence, for a particular experimental unit, the product of values of those two multiplicative components is equal to the yield value. Further on, consider root beet yield which is equal to the sum of its additive yield components, namely yields of roots from different root fractions.

Most agronomy and plant breeding investigations deal with multiplicative yield components. Thus in this paper we consider only multiplicative yield component analysis; for the sake of simplicity, we will call it yield component analysis (YCA), whereas multiplicative yield components will be called yield components. Note that the definition of yield components provided in the preceding paragraph is not limited to most common yield components that are usually considered for particular crop species. For this reason, yield component analysis, understood as a statistical methodology, is applicable for any system in which a response variable is to be considered as an effect of several dependent variables, which values multiplied are equal to the response variable value (for a particular experimental unit as well as for the statistical population).

YCA is a crucial approach in many agronomical and plant breeding studies, mostly because yield components provide a phenotypic synthesis of the ontogenetic cycle in a process of yield formation. It should allow clarifying how variation in yield components, resulting from genetic, environmental, and management factors, affects variation in crop yield. Then, this procedure should give information how yield components influence yield formation at the genetic, environmental, or phenotypic level. Moreover, YCA should be powerful in quantification of a level at which each component accounts for the yield

DOI: 10.1556/CRC.34.2006.2-3.222

variability; in such a case it would provide a relative contribution of each component to the final yield variability. Knowledge of the nature of such a phenomenon is important in many research areas like crop physiology, quantitative genetics, plant breeding, and crop management. It would enable to (i) provide a detailed overall view on a process of final yield formation (Board et al., 2003; García del Moral et al., 2003; Mohammadi et al., 2003); (ii) identify those yield components which relatively strongly affect final yield; the components identified as important in affecting yield can be used as selection criteria (for indirect selection or selection indices) in choosing genotypes having high yield potential in some environments (Board et al., 1997, 2003; Denčić et al., 2000); and (iii) reveal those components that are critical in controlling crop yield of the registered varieties grown across a target region environments (Board et al., 1999; Jag et al., 2000; Maman et al., 2004).

Various methods and techniques have been proposed and applied for YCA. Multiple linear regression and its modifications has been most exploited; these modifications are path analysis (e.g., Board et al., 1997, 1999, 2003; García del Moral et al., 2003; Mohammadi et al., 2003; Maman et al., 2004) and techniques based on stepwise regression, e.g., sequential yield component analysis SYCA (e.g., McArthur and Eaton, 1988; Gołaszewski, 1996) or its modifications (Sparnaaij and Bos, 1993; Mądry et al., 2005; Kozak and Mądry, 2005).

Studying multiplicity of papers, in which yield component analysis is either investigated from a methodological point of view or just applied for agronomy / plant breeding data, gives rise to put the following questions: Is a choice of a statistical approach to be applied in yield component analysis important? Can different statistical methods lead to different results and, in turn, interpretation of the phenomenon studied? Then, if so, which statistical method(s) of YCA should be applied? These questions are not easy to answer; actually, some of them still have no answer. Our aim in this paper is to try to answer those questions, if possible; otherwise, i.e., if the answer is not available, we discuss the problem and try to provide some hints and the direction for future work on it.

2. Modelling yield component analysis

Multiplicative yield model, which describes mathematically relationship between yield and its components, has a form

$$Y = f(X_i) = \prod_{i=1}^{k} X_i \qquad (1)$$

where Y stands for final crop yield per plant or unit area in its respective population under study, and X_i ($i=1,...,k$) are yield components.

Model (1) is deterministic, not stochastic, because it does not include any residual (error) term (Kozak, 2004). In other words, this model explains yield variability in a population of studied experimental units completely (Sparnaaij and Bos, 1993). In order to express the multiplicative pattern of yield, the model (1) should be fulfilled both for the population and for sample measurements of yield and its components. (Here it is assumed that the experimental units are randomly chosen as a representative sample from the population under study.)

The yield model (1) can be illustrated with some typical examples. Winter rape seed yield per unit area is a product of two components: plant density (number of plants per unit area) and average yield per plant (Hűhn, 1987). Winter triticale grain yield per unit area can be treated as a product of two components: biomass and harvest index, as well as a product of three components: ear density (number of ears per unit area), average number of kernels per ear and average kernel weight (Giunta et al., 1999). Alfalfa seed yield per unit area can be considered as a product of four components: stem density (number of stems per unit area), average number of pods per stem, average number of seeds per pod, and average seed weight (Iannucci et al., 2002).

Note that the model (1) can be used to analyze any causal system for which the model is applicable, not only for yield and its classical components. Therefore, for instance, harvest index and biomass yield can be treated as grain yield components, because the product of their values for each experimental unit is equal to the grain yield value. This conclusion, which explicitly follows from methodological papers on YCA (e.g., Sparnaaij and Bos, 1993; Piepho, 1995; Kozak, 2004), is of great importance, because the yield component methodology may be applied in various investigations, even in which the response variable is not yield but some other characteristic. For example, we can employ this methodology to investigate affecting nitrogen uptake (the response variable) by its two components: nitrogen uptake efficiency and soil plus fertilizer nitrogen content (note that by multiplying the values of those two components we obtain the value of nitrogen uptake).

However, a common mistake, which is unacceptable, is to call each yield-contributing character a "yield component." Subject to their definition, yield components are such characters that their product is equal to yield (Kozak, 2004, and the previous citations). When analyzing the influence of many yield-contributing characters, one should always remember to distinguish yield components from other yield-contributing characters.

In a classical approach to YCA, the relationship (1) is approximated with the linear multiple regression model (Kozak, 2004):

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon , \tag{2}$$

where Y is final yield, and X_i (i=1,...,k) are its components, β_0 is the regression intercept, β_i are the partial regression coefficients, ε is the residual. The model (2) is usually considered in its standardized form

$$y = \sum_{i=1}^{k} p_i x_i + \varepsilon , \tag{3}$$

where y and x_i are standardized Y and X_i, respectively (i.e., they all have mean 0 and variance 1), and p_i are the standardized partial regression coefficients, commonly called the path coefficients. The model (3) is a basic model for path analysis, the method most often used in YCA.

Although applying the linear model (2) or (3) in multiplicative yield component analysis is, in fact, time-honoured and common approach, it is disputable. Let us compare the multiplicative model (1) with the linear model (2) or (3). Applying the regression model (2) for describing the multiplicative relationship (1) always provides some errors (ε in the models (2) and (3)) which result only from inappropriateness of the linear model in modelling the studied relationship (1). Thus, the errors ε result from bad fitting of the linear model to the multiplicative one, so the classical assumption of regression analysis (normal distribution of errors) is not fulfilled. In spite of this, the model (2) can be regarded as an approximation of the multiplicative relationship (1); such an approach is commonly practiced in agronomy and plant breeding investigations. Although in many studies a multiple linear determination coefficient (R-square) is large (bigger than 90%), what could be an empirical justification of such an approach, in a recent paper by Kozak (2004) it has been shown that it is possible to obtain the R-square less than 70-80%, and even close to 0%. It depends mostly on the variability in variables under consideration, i.e., yield and its components.

Thus, the multiplicative YCA based on the linear model is model-biased. If the linear approximation of the model (1) is sufficiently good (i.e., the R-square value is reasonably large), YCA based on that model may provide appropriate interpretation of the phenomena studied, (although we are never sure of that). Otherwise, i.e., if the R-square is relatively small, the analysis would provide erroneous interpretation and conclusions.

Because of the nature of the relationship described by the multiplicative model (1), yield and its components should be measured in an appropriate way. For each experimental unit, the product of yield components measurements should be equal to yield measured (Kozak and Mądry, 2004). Only in such a case the model (1) is applicable at the sample level. The best way to fulfill this condition is to measure, at harvest, all characters, i.e., yield and its components, on entire experimental units (plots or sub-plots) belonging to a sample. This is, in fact, probably not possible in most investigations. Hence another appropriate approach is to observe $k-1$ yield components on entire experimental units (or on the same part of the units), and to calculate values of one component (the one which is the most difficult to measure, or the one usually measured with the biggest error; for instance, in cereals it would be average number of kernels per ear), say X_j, using the formula (Kozak and Mądry, 2004):

$$X_{jl} = \frac{Y_l}{\prod_{i=1, i \neq j}^{k} X_{il}} \tag{4}$$

where X_{jl} is the calculated value of the j-th ($j=1,...,k$) yield component for the l-th ($l=1,...,n$, n being the sample size) experimental unit, Y_l and X_{il} ($i=1,...,k$, $i \neq j$) are observed values of yield and i-th yield component for the l-th experimental unit, respectively.

For cereals, such direct and indirect measuring of the set of the characters (Y, X_1, ..., X_k) has been practised sometimes, but quite often scientists have observed yield on entire plots, whereas average number of kernels per ear (panicle) as well as average kernel weight have observed on a sub-sample of ten or thirty ears or panicles collected randomly from a plot (e.g., Giunta et al., 1999; Denčić et al., 2000; Sinebo, 2002; García del Moral et al., 2003, Maman et al., 2004). Number of ears per unit area has been counted usually on a sub-sample of a plot. Donaldson et al. (2001) collected data for yield components using the technique described by the formula (4), but they observed yield on another part of a plot.

Such measuring of yield and its components, as in the cited papers, usually provides substantial sample errors of yield components observation. These sample errors can cause inference in YCA inappropriate. Magnitude of these errors may be impressive, since values of the R-square obtained in the cited studies were usually quite low.

3. Sequential or nonsequential approach to analysis?

There is another important question and problem in YCA: Should one apply a sequential approach to YCA (e.g., Sparnaaij and Bos, 1993; García del Moral et al., 2003) or nonsequential one (e.g., Hühn, 1987; Piepho, 1995)? There has been a large discussion on this problem, and from it follows a simple conclusion and methodological rule: If one consider yield components that occur (develop) successively (sequentially) during plant ontogeny, a sequential approach to YCA should be applied (Sparnaaij and Bos, 1993; García del Moral et al., 2003, Mądry et al. 2005). A special pattern of relationship among sequentially developing components is as follows: a particular yield component may be affected by yield component which developed earlier, but the former cannot affect the latter. Hence first yield component may be a cause of all other components, but it is not their effect. The last component may be the effect of all other components involved in the yield model (1), but it cannot affect them.

On the other hand, if we consider yield components which finish their development at the same time during plant ontogeny (i.e., components being co-related), we should apply a nonsequential approach to YCA (Kozak, 2004; Kozak and Mądry 2005). Earlier some scientists were claiming that yield components develop sequentially (cf., e.g., Sparnaaij and Bos, 1993), whereas others were asserting that components develop nonsequentially (cf., e.g., Piepho, 1995). Authors of each such a paper presented some examples which had been to confirm the truth of their statement.

In fact, some components develop (occur) at the sequential order (e.g., the simplest example of classical yield components, the number of plants per unit area and average plant yield), and others develop nonsequentially (e.g., biomass yield and harvest index). Well known cereal grain yield components (i.e., number of ears per m^2, average number of kernels per ear, and average kernel weight) develop at the sequential order (García del Moral et al., 2003).

Identification of a real pattern of a relationship between yield and its components (i.e., decision if it is a sequential or nonsequential development) is crucial for proper description and interpretation of this relationship. For this reason it should be an important criterion of a choice of a statistical method to be applied. One should choose the method which is the most close in its mathematical philosophy to reality of the phenomena, with respect to sequential or nonsequential pattern of the development of the traits.

4. A choice of a statistical method for yield component analysis

When discussing a choice of a statistical method to perform YCA, we will take account of all methodological issues discussed in previous sections. Therefore, a statistical method for YCA should take into account (i) multiplicative model of a relation under consideration, both in a population of experimental units and in a sample from it; and (ii) a sequential or nonsequential pattern of development of yield components. Such a method should also provide a 100% determination coefficient for the estimated model of yield versus its components.

For this reason, we need two statistical methods for YCA: one for the sequential, and second for the nonsequential case. We will not present here any detailed discussion on the usefulness of methods commonly used in YCA. We will not also give any statistical details of the methods. The reader can find them in the cited papers and, what is the most important, he can judge them subjectively, with respect to the above notes given in the paper.

It seems that a method proposed and described by Kozak (2004) can be seen as appropriate for the nonsequential case. This method is not based on the linear model (2) but on the multiplicative model (1). The coefficients of the influence of yield components on yield, which are based on the model (1), are quite similar in interpretation to the path coefficients. Unfortunately, this method cannot be applied for the sequential case. Moreover, in our opinion none of the proposed methods for YCA is indisputably appropriate in the sequential case. The method of Sparnaaij and Bos (1993), together with its modification presented by Kozak and Mądry (2005) and applied by Samborski et al. (2005), could be seen as the best method at the moment. But we think that this problem still requires more work on.

5. Conclusion

The issues we have paid attention to in sections 2, 3, and 4 have quite an impact on proper interpretation of the relationship among yield and its components. If one makes incorrect choices connected with any of these issues, he will be likely to interpret the phenomenon studied improperly. Thus it is of great importance to stimulate further discussion on the methodology of yield component analysis, since existing methods of YCA are still not satisfying. This paper is to be a next step in this discussion, which, we hope, provides some vision for further studies.

Unfortunately, it turns out that most applications of yield component analysis are not correct, because some of the choices made by researchers are disputable what, in turn, resulted in low value of R-square. For instance, Maman et al. (2004) obtained 41% determination of pearl millet grain yield modelled by its components: panicle per m^2, kernel per panicle, and kernel weight. Gan et al. (2003) studied a relationship between seed yield of chickpea and its components (pods per m^2, seeds per pod, and seed weight); they obtained the following determination coefficients: 42.5% (Desi-chickpea on conventional summerfallow), 18.0% (Desi-chickpea on no-till wheat stubble), 59.6% (Kabuli-chickpea on conventional summerfallow), and 39.9% (Kabuli-chickpea on no-till wheat stubble). Generally speaking, a value of determination coefficients at the level of 80-90% is common. As we have mentioned, such values of the determination coefficient are unacceptable in yield component analysis. It is also of great importance to choose an appropriate approach to the analysis: the sequential or nonsequential one. A bad choice would lead to incorrect results of YCA and, in turn, false interpretation. This issue may be perfectly understood by studying the results presented by Piepho (1995) who applied nonsequential approach to components which develop sequentially. The results were so much different that it is obvious that one cannot just toss a coin to help to decide what the choice should be.

Yield component analysis is not a classical statistical problem, and it needs a special approach to. Moreover, do not try to think about yield components in the same manner as you do in a case of other yield-contributing characters. It is a good way to misinterpret a pattern of affecting yield by these particular crop and plant characters – yield components.

This paper is not an exposition of yield component analysis. Our objective was to direct your thoughts at the problem rather than to give its detailed description; the details are presented in the papers that we have refereed to. We hope that this short note will stimulate a discussion on the methodology of yield component analysis as well as on its appropriate application.

References

1. Board, J.E., Kang, M.S. and Bodrero, M.L. (2003). Yield components as indirect selection criteria for late planted soybean cultivars. *Agron. J.* 95, 420-429.
2. Board, J.E., Kang M.S., Harville, B.G. (1997). Path analyses identify indirect selection criteria for yield of late-planted soybean. *Crop Sci.* 37, 879–884.
3. Board, J.E., Kang M.S., Harville, B.G. (1999). Path analyses of the yield formation process for late-planted soybean. *Agron J.* 91, 128-135.

4. Denčić, S., Kastori, R., Kobiljski, B., Duggan, B. (2000). Evaluation of grain yield and its components in wheat cultivars and landraces under near optimal and drought conditions. *Euphytica* 113, 43-52.
5. Donaldson, E., Schillinger, W.F, Dofing, S.M. (2001). Straw production and grain yield relationships in winter wheat. *Crop Sci.* 41:100-106.
6. Gan, Y.T., Liu, P.H., Stevenson, F.C., McDonald, C.L. (2003). Interrelationships among yield components of chickpea in semiarid environments. *Can. J. Plant Sci.* 83, 759-767.
7. García del Moral, L.F., Rharrabti, Y., Villegas, D. and Royo, C. (2003). Evaluation of grain yield and its components in durum wheat under Mediterranean conditions: An ontogenetic approach. *Agron. J.* 95, 266-274.
8. Giunta, F., Motzo, R., Deidda, M. (1999). Grain yield analysis of a triticale (*XTriticosecale Wittmack*) collection grown in a Mediterranean environment. *Field Crops Res.* 63, 199-210.
9. Gołaszewski, J. (1996). A method of yield component analysis. *Biometrical Letters* 33, 79-88.
10. Hühn, M. (1987). Stability Analysis of Winter-rape (*Brassica napus* L.) by Using Plant Density and Mean Yield per Plant. *J. Agron. & Crop Sci.* 159, 73-81.
11. Iannucci, A., Di Fonzo, N., Martiniello, P. (2002). Alfalfa (*Medicago sativa* L.) seed yield and quality under different forage management systems and irrigation treatments in a Mediterranean environment. *Field Crops Research 78, 65-74.*
12. Jag, S., Hariprasad, A.S., Lakshmi, K., Mani, V.P., Chauhan, V.S., Shoran, J., Kant, L. (2000). Association and contribution of yield attributes to seed yield in wheat under varying environments in north western hills. *Annals of Agricultural Research*, 21, 274-278.
13. Kozak, M. (2004). New concept of yield component analysis. *Biometrical Letters* 41(2), 59-69.
14. Kozak, M., Mądry, W. (2004). Statistical analysis of multiplicative crop yield components – background of yield modeling. *Postępy Nauk Rolniczych* 5, 13-25. (in Polish)
15. Kozak, M., Mądry, W. (2005). Primary characters in yield component analysis. Part I. Description of the method. *Fragmenta Agronomica* 4 (in press, in Polish).
16. Maman, N., Mason S.C., Lyon, D.J., Prabhakar, D. (2004). Yield components of pearl millet and grain sorghum across environments in the Central Great Plains. *Crop Sci.* 44, 2138-2145.
17. Mądry, W., Kozak, M., Pluta, S., Żurawicz, E. (2005). A new approach to sequential yield component analysis (SYCA): application to fruit yield in blackcurrant (*Ribes nigrum* L.) *J. New Seeds* 7: 85–107.
18. McArthur, D.A.J, Eaton, G.W. (1988). Strawberry yield response to fertilizer, paclobutrazol and chlormequat. *Scientia Horticult.*, 34, s. 33-45.
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+ Sci.* 43:1690-1697.
20. Piepho, H.P. (1995). A simple procedure for yield component analysis. *Euphytica* 84, 43-48.

21. Samborski, S., Kozak, M., Mądry, W., Rozbicki, J. (2005). Primary characters in yield component analysis. Part II. Application for winter triticale grain yield. *Fragmenta Agronomica* 4 (in press, in Polish).
22. Sinebo, W. (2002). Yield relationships of barleys grown in a tropical highland environment. *Crop Sci.* 42: 428-437.
23. Sparnaaij, L.D., Bos, I. (1993). Component analysis of complex characters in plant breeding. I. Proposed method for quantifying the relative contribution of individual components to variation of the complex character. *Euphytica* 70, 225-235.

Received 18 November, 2005, accepted 24 March, 2006