CBCS
http://agrobiol.sggw.waw.pl/cbcs

Communications in Biometry and Crop Science
Vol. 5, No. 2, 2010, pp. 83–95

International Journal of the Faculty of Agriculture and Biology,
Warsaw University of Life Sciences, Poland

REGULAR ARTICLE

Use of parallel coordinate plots in multi-response selection of interesting genotypes

Marcin Kozak

Department of Experimental Design and Bioinformatics, Warsaw University of Life Sciences,
Nowoursynowska 159, 02-776 Warsaw, Poland.
E-mail: nyggus@gmail.com

CITATION: Kozak, M., (2010). Use of parallel coordinate plots in multi-response selection of interesting genotypes. *Communications in Biometry and Crop Science* 5 (2), 83–95.

Received: 25 March 2010, Accepted: 4 September 2010, Published online: 19 September 2010

ABSTRACT

Visualizing genotype-by-environment interaction in the case of several attributes, a situation that is often dealt with in plant breeding, is discussed in the paper. The parallel coordinate plot is proposed as an efficient tool for such visualization. Various applications of this type of graph are presented for across-environment and environment-wise plotting, thereby offering rich information about genotypes' performance. It is shown how one can analyze the data and report the results by means of the parallel coordinate plots.

Key Words: *adaptability; genotype selection; genotype-by-environment interaction; visualization.*

INTRODUCTION

Parallel coordinate plots (PCPs) are an efficient tool for visualizing multivariate data (Inselberg, 1985; Wegman, 1990). They have been used in various research areas, although environmental and agricultural applications are rather scarce; for example, Andrienko and Andrienko (2001) applied PCPs for exploring spatial data, and Villamil et al. (2008) for analyzing soil quality data. Taking into account what interpretation possibilities this type of plot offers, it may be useful in multivariate genotype selection.

Genotype selection in the case of multiresponse data is not an easy matter. Regular statistical analyses provide a basis for such selection, but visualization may offer additional, very powerful tools. Univariate data (e.g., yield of genotypes in one environment) are easy to analyze, and the best genotypes are easy to select. However, with additional dimensions, for example many environments and many attributes, such data become very difficult to study. For such selection, multivariate selection indices might be used (Brown, 1988), but any index, as a single number, cannot provide detailed insights into the genotypes' performance—when choosing promising genotypes for inclusion in breeding programs, plant breeders choose the

best genotypes as well as those which may introduce some interesting attributes to the breeding pool. The latter group of genotypes might be lost if one would base only on multivariate selection indices (although the indices can be used with various criteria, in that way providing more numbers than just one—so again, the data become complex to analyze).

For bi-dimensional data (genotypes × attributes), biplot-based approaches can be applied. In the case of three-dimensional data (genotypes × environments × attributes), combinations of bivariate plots can be used (e.g., Manson et al., 2008) to support interpretation and genotype selection. Three-mode principal component analysis (Kroonenberg and Basford, 1989) supported by a so-called joint biplot has been used to study the relationships among genotypes, environments and attributes (e.g., Bertero et al., 2004; Varela et al., 2006; D'Andrea et al., 2008). However, interpretation of the three-mode PCA is not easy. Kozak et al. (2008) proposed a multivariate selection of promising genotypes based on joint cluster and path analyses, but this approach does not facilitate seeing the genotypes' performance in terms of all the attributes. Gupta et al. (2009) showed how intuitive selection can be made in the case of three-dimensional data attribute × genotype × year; this method nevertheless pictures only bi-attribute performance of the genotypes, not accounting for multi-attribute ones. Basford and Tukey (1999) describe techniques for graphing multiresponse data, most of which are based on performance plots and scatterplot matrices.

In this paper it is shown that a parallel coordinate plot can be a very useful tool for selection of interesting genotypes in the case of many attributes of interest. What an "interesting genotype" is in the case of multiresponse data depends on how a breeder understands it and what the breeding program is aimed at; in fact, PCPs give space for various such understandings and aims. PCPs can be used at the analysis stage of the breeding research, which aims to select promising genotypes, and at the presentation stage, which aims to communicate the results of the analysis.

MATERIALS AND METHODS

PLANT MATERIAL

The data come from a two-year plant breeding experiment on soybean, conducted in four locations (Mungomery et al., 1974), which are given in Basford and Tukey (1999). For the present paper, 15 early genotypes (labeled with numbers 44-58) were chosen, studied in eight environments being the combinations of locations Brookstead, Lawes, Nambour and Redland Bay (Australia) and years 1970 and 1971. We will consider the following attributes: plant height (m), lodging (%), oil (%), protein (%), quality index (oil + protein, both in %), seed size (g/100), seed yield (t/ha), protein yield (t/ha), oil yield (t/ha), and economic yield (protein + oil yield, in t/ha). The value for a particular genotype in an environment is the mean value from two blocks. Basford and Tukey (1999) discuss the experiment in detail.

PARALLEL COORDINATE PLOT

The construction of a PCP is easy. Each attribute is represented on the x-axis and has its own y-axis; the y-axes are parallel, have the same length, and start with a minimum of the corresponding attribute and end with its maximum. (In fact, quite often the plot is rotated so that the attributes are presented on the vertical axis.) Note thus that when a particular genotype is placed in the middle of a y-axis, it does not mean that its value is around the mean of the corresponding attribute—it is the middle point within the attribute's range. For a particular genotype, the points on the adjacent y-axes are joined by a line, thereby picturing a multidimensional characterization (profile) of the genotype. Since many genotypes are plotted on the same PCP, a particular genotype's performance can be seen against a background of the whole pool of genotypes studied. One can (but does not have to) include in the plot the minimum and maximum values of the attributes.

An important issue concerned with PCPs is the order of attributes on the x-axis. As Huh and Park (2008) indicate, the relationships between variables in non-adjacent positions are

difficult to figure out. One can draw all possible PCPs for a data set (it can be quite a large number—see Wegman, 1990) and choose those that offer the most interesting information; alternatively, one can apply some ordering of the variables (e.g., Hurley, 2004). However, note that these rules aim to support discovering multivariate relations among the variables, while in our problem the aim is to select promising genotypes and interpret their performance against the background of all genotypes. Thus several other rules might be considered useful in choosing this order for the genotype selection. The last attribute should be the most economically important one (e.g., oil, protein or economic yield in our example), if there is such. Attributes that are somewhat related or the relation of which may be interesting (e.g., oil and protein content, depending on the analysis's aim) should be placed adjacently. If possible, attributes may be ordered by the order of their development during plant ontogeny (see, e.g., Mądry et al., 2005). If there are several orderings that could be followed according to the above rules, one should try all of them and choose those which offer the most interesting information. What one considers interesting depends on the aim of the selection—but to make PCPs useful, all attributes used for plotting should be interesting. If an attribute is not used in interpretation at all, it should be removed from the plots.

Such a general interpretation of data (Wegman, 1990), which may include grouping of genotypes and correlations among attributes, should constitute a first stage of the analysis. For multi-environment data, the PCP should be constructed for means of the attributes across environments. This provides the overall (across-environment) profiles for the genotypes. This interpretation should be supported by other tools for interpreting multivariate data, for example the scatterplot matrix (Cleveland, 1994). PCPs based on attribute means from environments will be hereafter called the across-environment PCPs.

Next one can look for interesting genotypes by constructing an across-environment PCP for each genotype in such a way that this genotype is drawn with a black line while the remaining genotypes are drawn with a grey line. Hence there will be as many plots as there are genotypes. In that way one can search for genotypes with some interesting features in a multivariate sense (e.g., these can be high oil and protein content, providing high quality index, with small lodging). We need to remember at this stage that the across-environment PCPs ignore the genotype-by-environment interaction that is very likely to be present in multiresponse multi-environment data. Still they can provide a general picture of genotypes, which is an important aspect of genotype selection.

A next stage, thus, should take into account the genotype-by-environment interaction, if it is indeed present in the data. For each genotype there will be as many PCPs as there are environments; such PCPs will be hereafter called the environment-wise PCPs. Thus at this stage there will be $E \times G$ PCPs, E being the number of environments and G the number of genotypes. This interpretation can be supported with a scatterplot matrix constructed for each environment. The plots for each environment should be ordered by an increasing value of one of the attributes (the most interesting one, economic yield in our example); this ordering should be made from left to right and from bottom to top (following the standard ordering in trellis displays—see Cleveland, 1994). Of course, this stage is very demanding for an analyst, but intensive interpretation cannot be avoided when there are many attributes, many genotypes and many environments. However, one can limit this stage to those genotypes that were selected as interesting based on across-environment PCPs. It is important to note that for a given genotype, a PCP for each environment is presented for environment-wise ranges of the attributes. Hence the environment-wise PCPs are not comparable among environments, although of course environment-wise PCPs for two genotypes in one environment are comparable.

The last stage is presentation of selected genotypes. An across-environment PCP for this still should include coordinates for all genotypes, but this time those selected as promising are drawn in black while those non-selected in grey. If there are too many selected genotypes to include in one PCP to make them easily distinguishable, they can be drawn on two (or

more) adjacent PCPs; some grouping may be applied here so that each PCP contains emphasized genotypes that are characterized by particular features (the background grey lines should present all genotypes except those emphasized on this PCP). If needed, such an across-environment PCP can be supported by chosen environment-wise PCPs.

All plots were drawn with R (R Development Core Team, 2009): the parallel coordinate plots using the `parcoord()` function of the library MASS (Venables and Ripley, 2002), while the scatterplot matrix using the `splom()` function of the library `lattice` (Sarkar, 2009).

RESULTS

Figure 1 shows an across-environment PCP for the 15 genotypes and all attributes of interest, while Figure 2 a scatterplot matrix for the same data. In Figure 1 minimum and maximum values are included for all the attributes; note that these values are minimum and maximum genotype means over environments (so they do not describe the minimum or maximum attribute values that were obtained in the experiments). Note also that the scatterplot matrix constructed for each environment might differ from that in Figure 2. From these figures it seems that most attributes were diverse, but worth noting is quite small diversity of quality index—it ranged between 59.6 and 63.4, showing that the genotypes were quite similar in terms of this attribute.

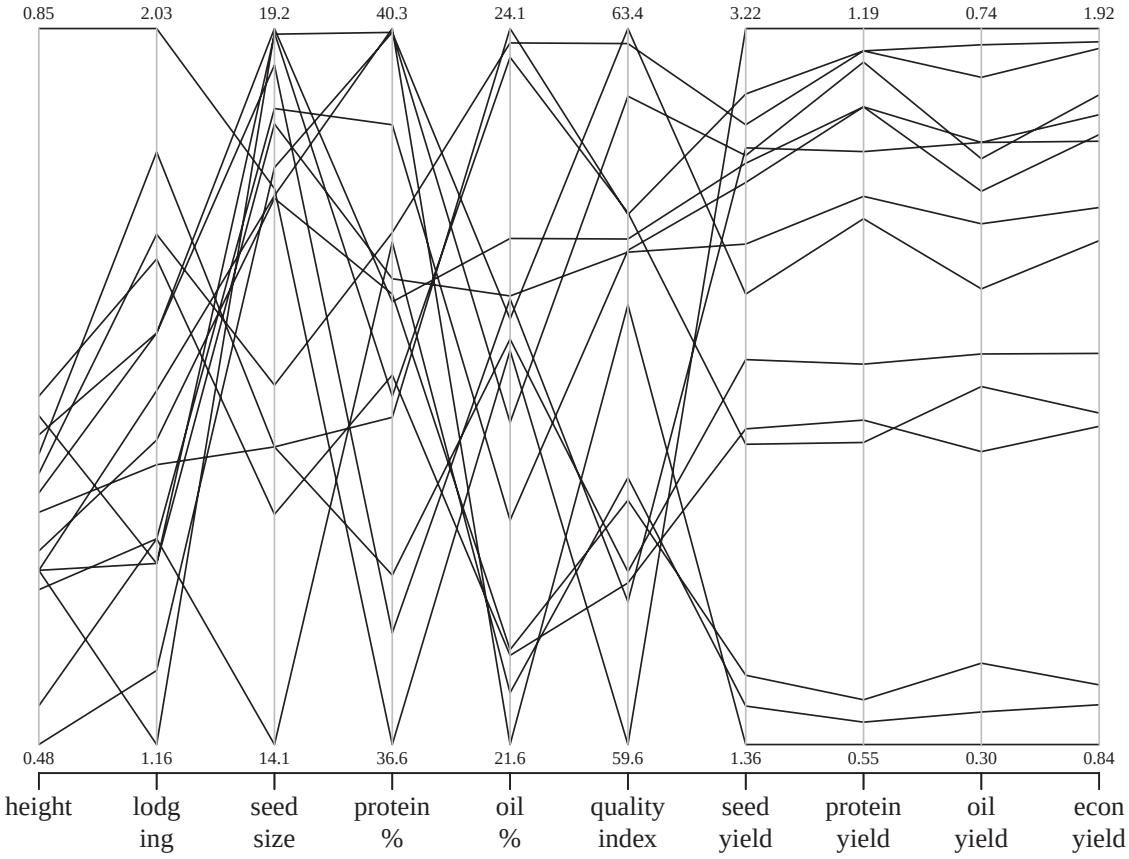

Figure 1. Parallel coordinate plot for 15 soy bean genotypes and all traits of interest.

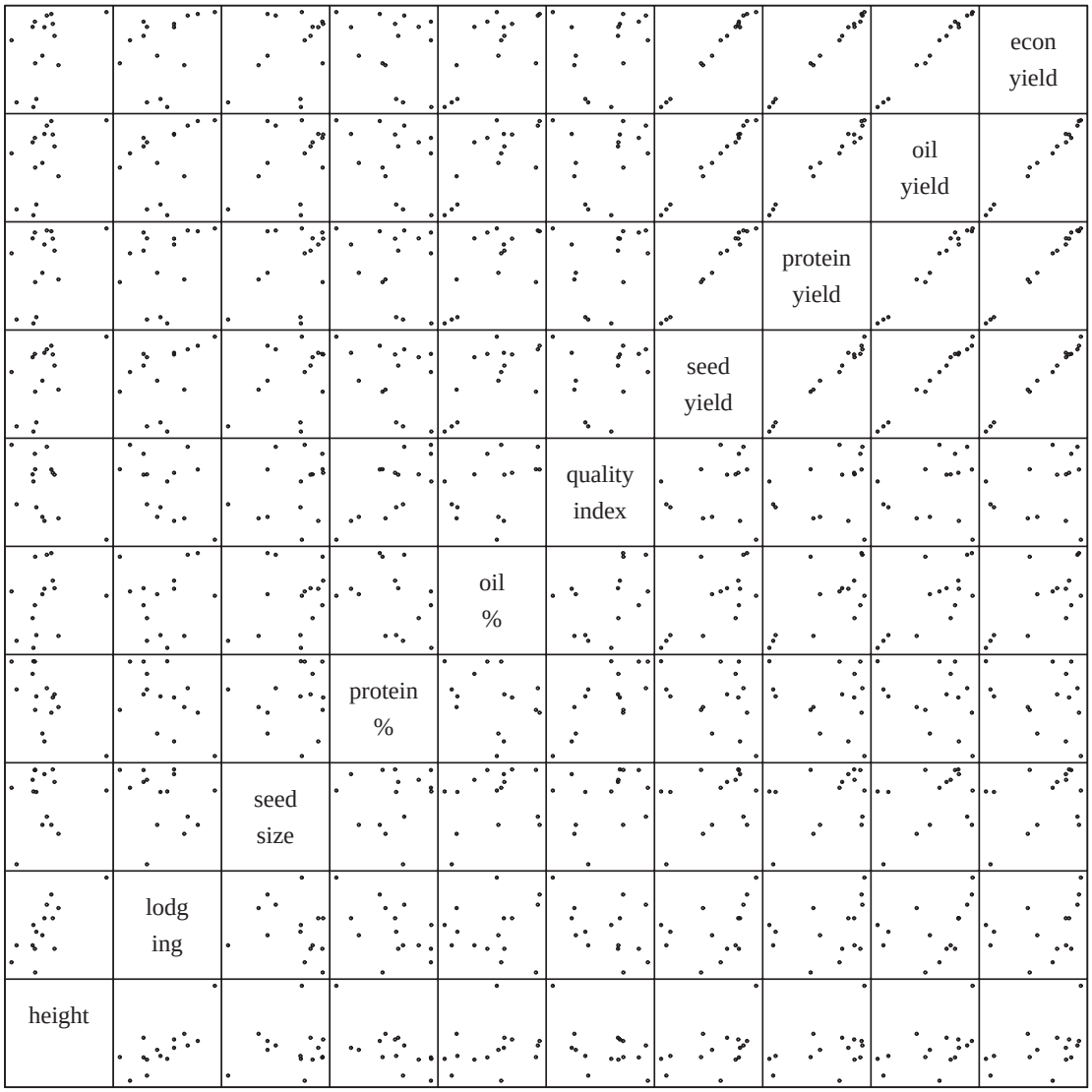

Figure 2. Scatterplot matrix for 15 soybean genotypes and all traits of interest. A point represents the means for the two corresponding traits across the eight environments.

Some other interesting general conclusions follow from these two plots. Protein, oil and economic yields were determined mainly by seed yield and were extremely strongly correlated (indeed, all Pearson's correlations among these four attributes are above 0.996). For this reason in genotype-wise plots we could leave out three of these four attributes. Since economic yield is the final attribute of interest, we could keep it, but in further analysis and interpretation we would have to keep in mind that seed yield is most important in determining economic yield, and quality index in fact seems to have no influence on it— which follows from Figure 2 (quality index does not influence economic yield while seed yield determines it almost totally). Nonetheless, for the purpose of the present paper it will be better to keep all the attributes in the plots.

Taller the plants tend to suffer more from lodging. A negative though rather weak relationship between oil and protein content was observed. Economic yield increased with an increase in oil content, seed yield, and protein and oil yield; it was related to neither protein content nor quality index. Lodging did not determine economic yield; in fact, genotypes with the highest seed yield and economic yield had the highest lodging.

After this general analysis, the second step follows, in which each genotype has its own across-environment PCP. See Figure 3 for the genotype-wise across-environment PCPs for genotypes 48, 49, 50 and 51 (for space reasons PCPs for all 15 genotypes are not presented). Note that here we need no maximum and minimum attribute values because they were already given in the previous PCP, and they would be repeated in each genotype-wise PCP. By looking at each such PCP, interesting insights into a particular genotype's performance follow. Genotype 48 had second the highest seed, protein, oil and economic yields, and the highest oil content. Even though genotype 49 had the tallest plants and greatest lodging, the lowest protein content and quality index (it is the right time to recall the small diversity of quality index) together with medium oil content, it had the highest seed, protein, oil and economic yields. Genotype 50 also performed very well. However, genotype 51 was the worst among the 15 genotypes in the study even though it had the highest protein content and medium quality index—the lowest seed yield accounted for the lowest protein, oil and economic yields. This genotype was chosen for presentation only to show its particular characteristics, not because of its high potential.

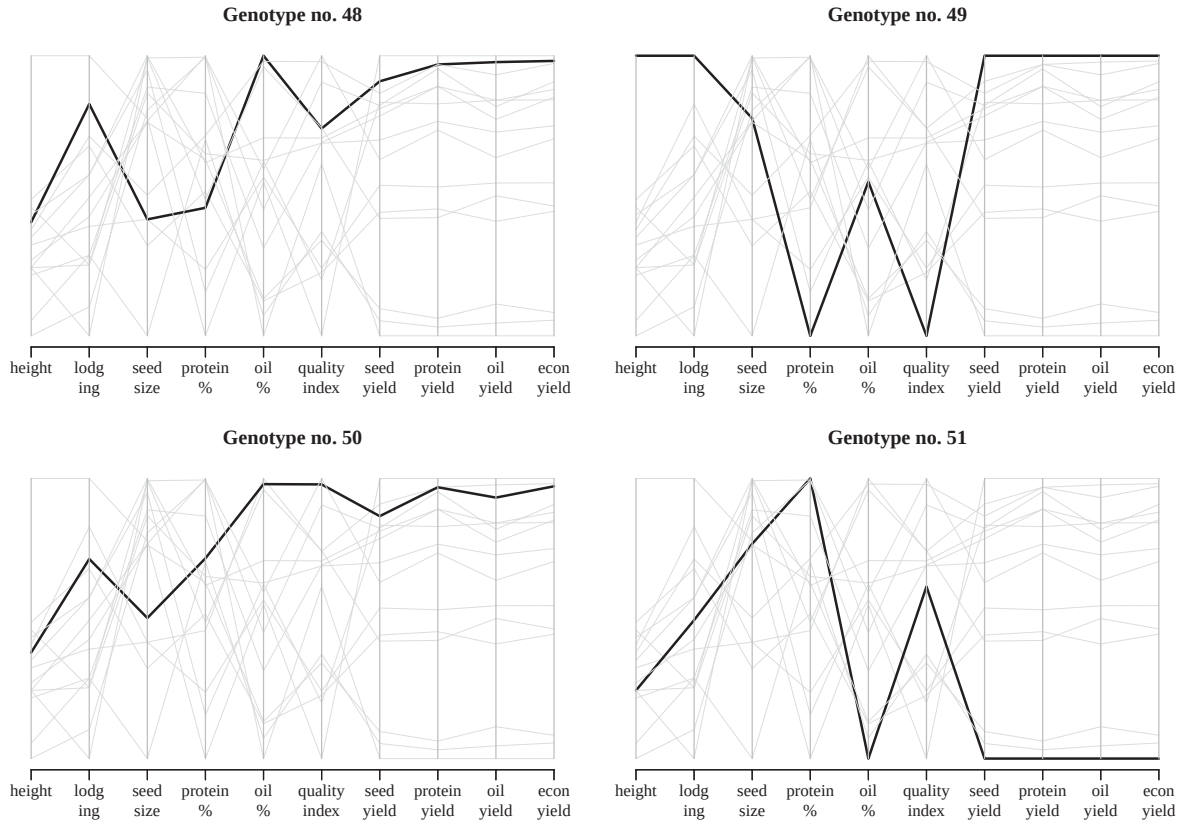

Figure 3. Example of genotype-wise parallel coordinate plots for four selected genotypes.

Figure 4 shows the three genotypes chosen above (so without genotype 51, which would not be considered interesting for further selection) against the background of the performance of all the genotypes studied. All that was seen for the selected genotypes on the genotype-wise plots can be now seen on this summary PCP; hence it is an optimum way of presenting the results of the across-environment analysis.

In the above examples the genotype-by-environment interaction was ignored. To take account of this interaction, one can plot the environment-wise PCPs for each selected genotype (or for each genotype). Three examples for genotypes 48, 49 and 50 are presented in Figures 5-7. Genotype 48 was very high yielding (in terms of all yields considered) in four

environments and medium to high yielding in other environments. Its lodging was from very small to high. Oil content was from medium to very high, while protein content was from low to medium. Genotype 49 had the tallest plants in all environments but one, and the highest lodging in several environments in which lodging was not the same for all genotypes. Its yield was very high in almost all environments. Protein content was low in all environments while oil content was medium in seven environments and the lowest in one environment. The strong genotype-by-environment interaction is revealed for genotype 50 in terms of oil and protein content. In some environments, this genotype had very high (even the highest) protein content, while in others medium or even the smallest; oil content was in some environments high while in others medium. Nonetheless, yield was high in each environment.

Note that for none of the three genotypes in Figures 5, 6 and 7 anything interesting concerned with the genotype-by-interaction could be observed thanks to the ordering of the plots subject to the increasing mean economic yield. In addition, for none of the 12 remaining genotypes this was observed. This is not, of course, the rule—sometimes certain genotypes will perform relatively very well only in poor environments while bad in good ones (or *vice versa*), which would indicate specific adaptation in terms of the corresponding attributes (Annicchiarico, 2002).

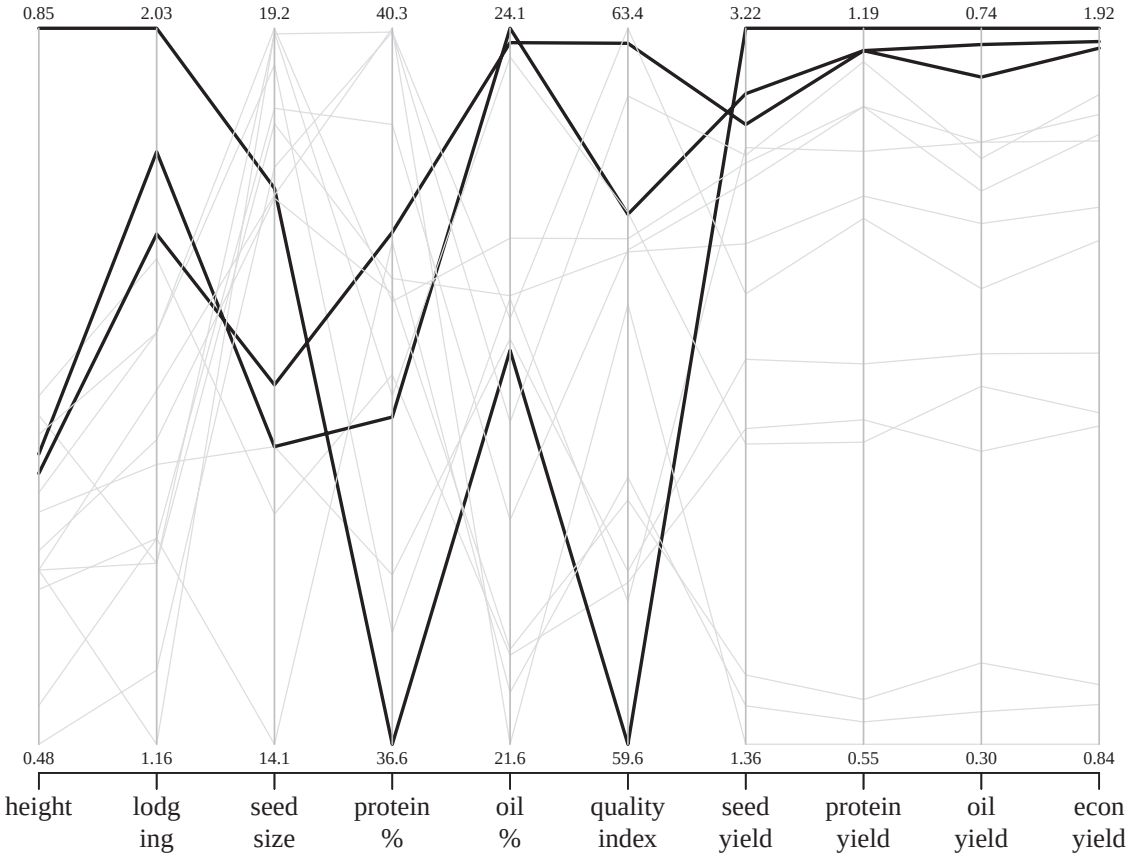

Figure 4. Parallel coordinate plot with genotypes 48, 49 and 50 emphasized by a thick black line.

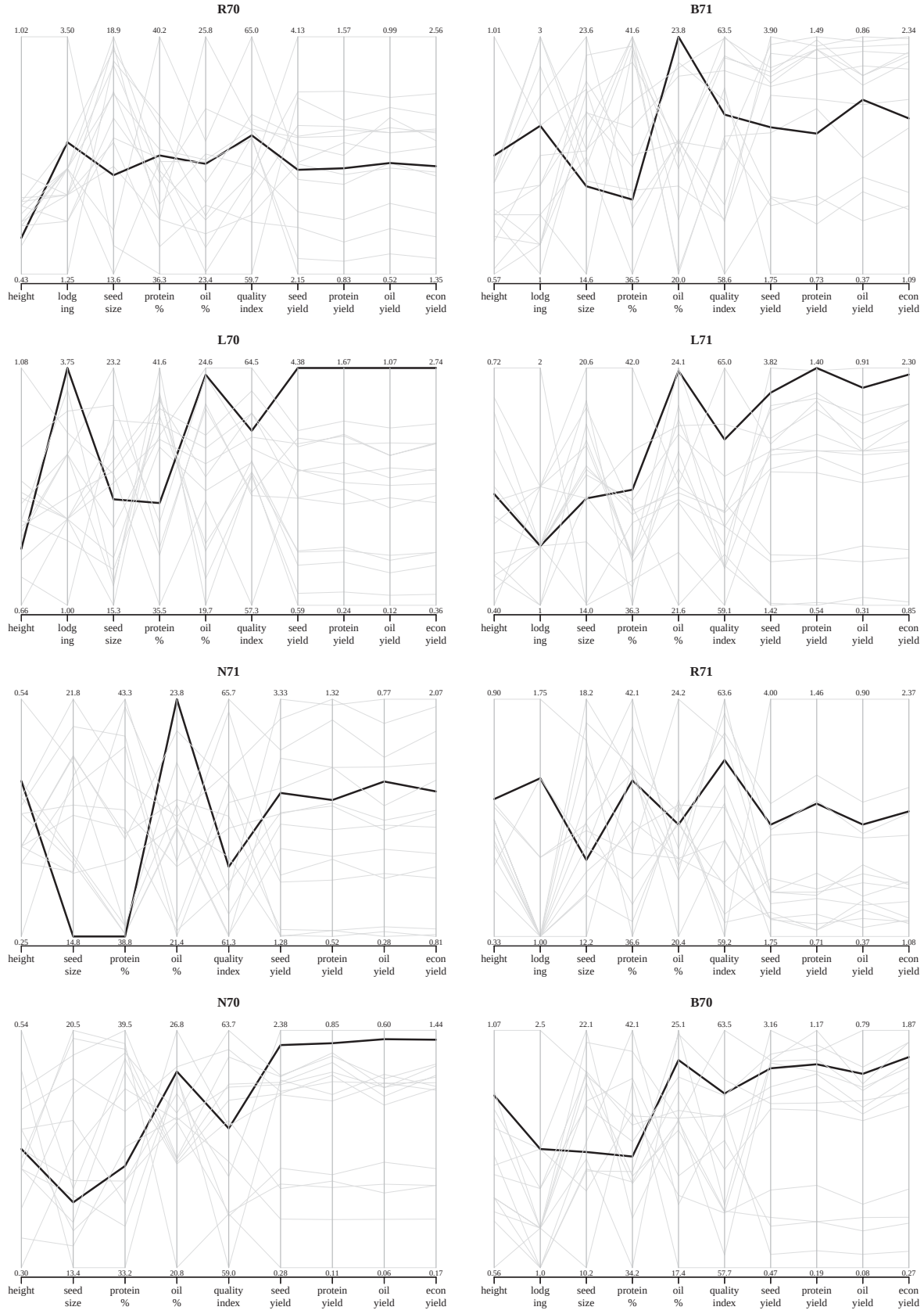

Figure 5. Parallel coordinate plots showing performance of genotype 48 in eight environments.

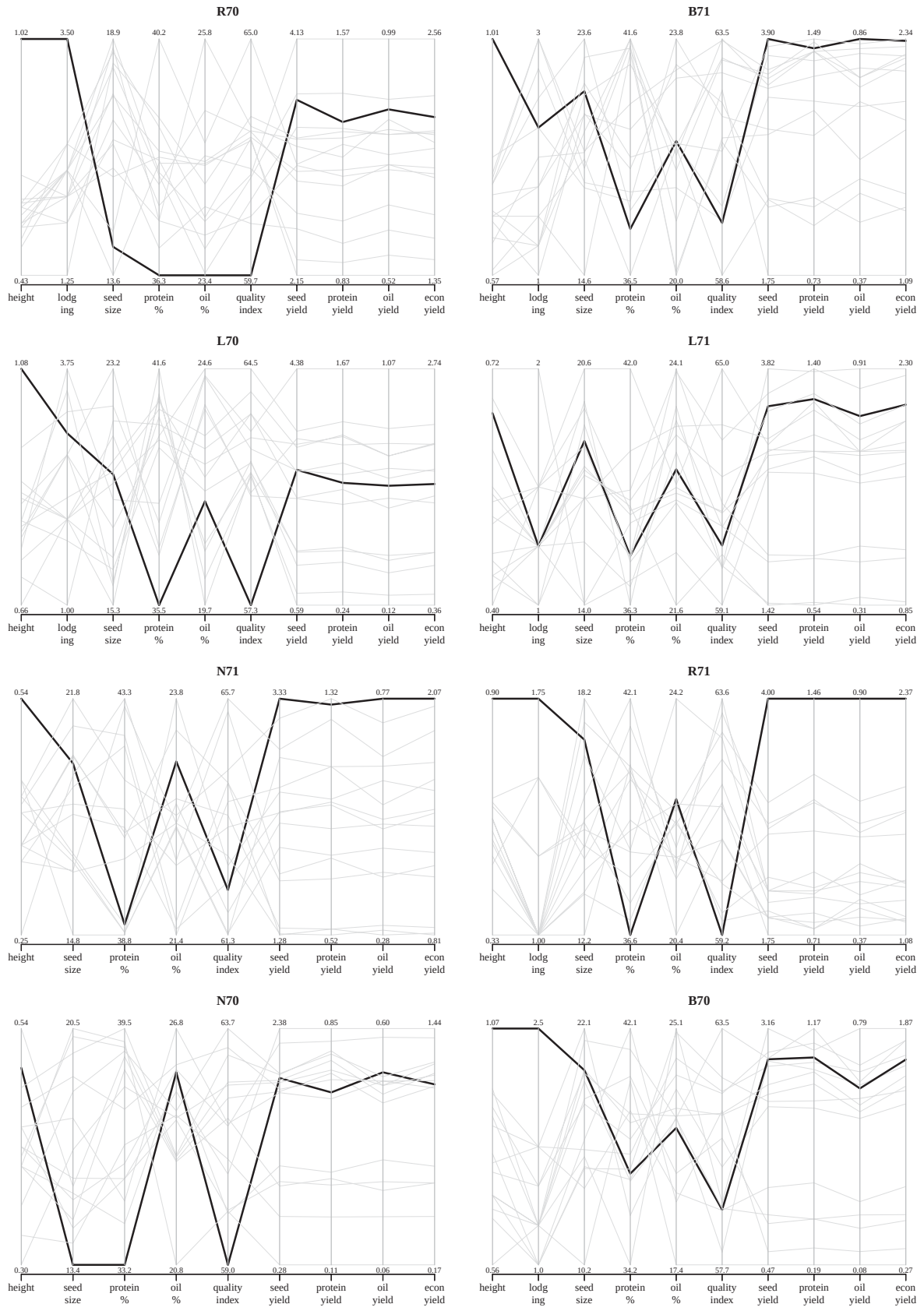

Figure 6. Parallel coordinate plots showing performance of genotype 49 in eight environments.

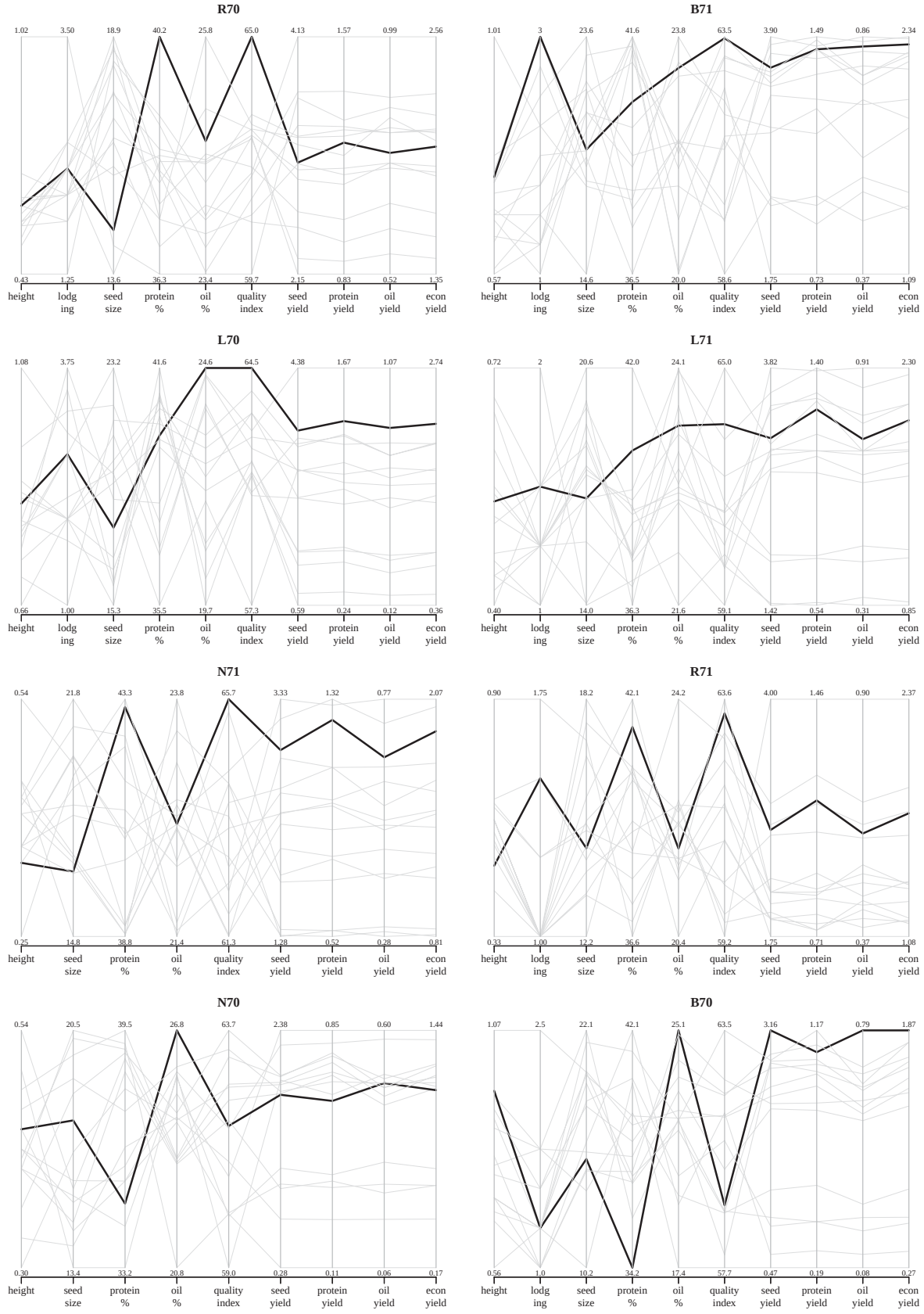

Figure 7. Parallel coordinate plots showing performance of genotype 50 in eight environments.

DISCUSSION

From the above examples it follows that parallel coordinate plots can be an efficient visualization tool for multivariate selection of promising genotypes. This efficiency has two main reasons. First, the genotype-wise PCPs offer quick access to information about performance of the genotypes. Second, this information is easily interpretable for plant breeders who do not possess advanced knowledge of statistics and visualization techniques. Although the example was based on 15 genotypes only, the PCP can be applied to any number of genotypes. This would of course require drawing and reading many graphs, but this cannot be avoided when one aims to choose promising genotypes from a large pool of them, studied in many environments. In most cases it would suffice to draw the environment-wise PCPs only for a number of genotypes selected based on the across-environment analysis. Nonetheless, this kind of analysis is limited to a reasonable number of environments and genotypes, not for technical but rather human reasons—it is not easy to understand patterns in data from a large number of genotypes with many traits. It would be difficult to analyze plots for say more than several environments; DeLacy et al. (1996) say that multi-environment trials in CIMMYT typically include 50 entries in over 60 environments, and for such data employing environment-wise PCPs seems impossible. However, note that PCPs can be applied also to visualize multiresponse profiles of genotypes already selected by means of other methods, including those for analyzing a single attribute (e.g., AMMI [Gauch, 1992] or GGE [Yan and Kang 2003]) and the pattern analysis based on the three-mode principal component analysis (Kroonenberg and Basford, 1989; Bertero et al., 2004; Varela et al., 2006; D'Andrea et al., 2008). In that way the number of genotypes to study can be decreased—still, however, the number of environment plays role. It might be possible to group environments that are similar in terms of the traits of interest, but this requires further studies. Three-mode PCA is not easy to conduct and interpret, and is not free of choices one has to make before the analysis based on the scope of the analysis (la Vega et al., 2002). This method can be very powerful in describing patterns within the data, yet the approach presented herein can provide a fairly easy (though time-consuming) way of interpretation of multi-attribute performance of particular genotypes.

The general interpretation of the data (like in Figure 1) is interesting, but this should be supported by other multivariate techniques (the scatterplot matrix was employed in the paper—Figure 2). This is especially important for choosing attributes to include in the genotype-wise PCPs. The main power of PCPs, however, is revealed at the analysis stage, by plotting the genotype-wise plots. This type of plotting offers clear information on how a particular genotype performs against the background of all genotypes, which is a very important piece of information for plant breeders; this is done for each environment as well as across environments. In fact, this could not likely be seen without graphing, and in addition, it seems that no commonly used graph can provide it in such detail. Finally, the PCP with emphasized genotypes selected in that way is also quite an efficient way of presenting the performance of these genotypes, although averaged across environments.

In summary, the parallel coordinate plots used as described in this paper can be an efficient tool in selection of promising genotypes. It describes across-environment and environment-wise multiresponse profiles of genotypes, supporting interpretation of their performance in terms of all traits of interest. Reading and interpreting the graphs is tedious, but this cannot be avoided with such complex data. The construction of graphs is also not easy, so future research should focus on producing user-friendly software or code that do that automatically for the user. Interactive plots are a tool that can be extremely helpful for the parallel coordinate plots described in this paper, by letting the users quickly approach the information they need (for example, by choosing a particular genotype or environment of interest).

REFERENCES

Andrienko, G., Andrienko, N. (2001). Exploring spatial data with dominant attribute map and parallel coordinates. *Computers, Environment and Urban Systems* 25, 5–15.

Annicchiarico, P. (2002). *Challenges and Opportunities for Plant Breeding and Cultivar Recommendations*. FAO Plant Production and Protection Papers 174.

Basford, K.E., Tukey, J.W. (1999). *Graphical analysis of multiresponse data. Illustrated with a plant breeding trial*. Chapman & Hall/CRC, Boca Raton.

Bertero, H.D., de la Vega, A.J., Correa, G., Jacobsen, S.E., Mujica, A. (2004). Genotype and genotype-by-environment interaction effects for grain yield and grain size of quinoa (*Chenopodium quinoa* Willd.) as revealed by pattern analysis of international multi-environment trials. *Field Crops Research* 89, 299–318.

Cleveland, W.S. (1994). *The elements of graphing data*. 2nd ed. Hobart Press, Summit, New Jersey, USA.

DeLacy, I.H., Basford, K.E., Cooper, M., Bull, J.K., McLaren, C.G. (1996). Analysis of multi-environment trials—an historical perspective. In: Cooper, M., Hammer, G.L. (Eds.). *Plant Adaptation and Crop Improvement*, Oxford: CAB International, 39–124.

de la Vega, A.J., Hall, A.J., Kroonenberg, P.M. (2002). Investigating the physiological bases of predictable and unpredictable genotype by environment interactions using three-mode pattern analysis. *Field Crops Research* 78, 165–183.

Gauch, H.G. (1992). *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*. Elsevier, New York, New York.

Gupta, S., Kozak, M., Sahay, G., Durai, A.A., Mitra, J., Verma, M.R., Pattanayaki, A., Thongbam, P.D., Dasi, A. (2009). Genetic parameters of selection and stability and identification of divergent parents for hybridization in rice bean (*Vigna umbellata* Thunb. (Ohwi and Ohashi)) in India. *Journal of Agricultural Science* 147, 581–588.

Huh, M.H., Park, D.Y. (2008). Enhancing parallel coordinates plots. *Journal of the Korean Statistical Society* 37, 129–133.

Hurley, C.B. (2004). Clustering visualizations of multidimensional data. *Journal of Computational and Graphical Statistics* 13, 788–806.

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer* 1, 69–91.

Kozak, M., Bocianowski, J., Rybiński, W. (2008). Selection of promising genotypes based on path and cluster analyses. *Journal of Agricultural Science* 146, 85–92.

Kroonenberg, P.M., Basford, K.E. (1989). An investigation of multi-attribute genotype response across environments using three-mode principal component analysis. *Euphytica* 44, 109–123.

Manson, H., Goonewardene, L., Spaner, D. (2008). Competitive traits and the stability of wheat cultivars in differing natural weed environments on the northern Canadian Prairies. *Journal of Agricultural Science* 146, 21–33.

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.). *Journal of New Seeds* 7, 85–107.

R Development Core Team (2009*). R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org [19 Oct 2009].

Sarkar, D. (2008). *Lattice. Multivariate Data Visualization with R*. Springer.

Varela, M., Crossa, J., Rane, J., Joshi, A.K., Trethowan, R. (2006). Analysis of a three-way interaction including multi-attributes. *Australian Journal of Agricultural Research* 57, 1185–1193.

Venables, W.N., Ripley, B.D. (2002). *Modern applied statistics with S*. Fourth Edition. Springer, New York.

Villamil, M.B., Miguez, F.E., Bollero, G.A. (2008). Multivariate analysis and visualization of soil quality data for no-till systems. *Journal of Environmental Quality* 37, 2063–2069.

Wegman, E.J. (1990). Hiperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association* 85, 664–675.

Yan, W., Kang, M.S. (2003). *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists*. CRC Press. Boca Raton, FL.