Biometrical Letters
Vol. 47 (2010), No. 1, 57-68

Call for more graphical elements in statistical teaching and consultancy

Marcin Kozak[1], Jan Bocianowski[2], Sylwia Sawkojć[3], Agnieszka Wnuk[1]

[1]Department of Experimental Design and Bioinformatics, Warsaw University of Life Sciences,
Nowoursynowska 159, 02-776 Warsaw, Poland, e-mail: nyggus@gmail.com
[2]Department of Mathematical and Statistical Methods, Poznań University of Life Sciences,
Wojska Polskiego 28, 60-637 Poznań, Poland
[3]Department of Soil Environment Sciences, Warsaw University of Life Sciences,
Nowoursynowska 159, 02-776 Warsaw, Poland

SUMMARY

This paper suggests that, all too often, graphical elements are discounted in statistical practice. Properly constructed graphs can greatly help understand data and statistical analysis, so the more of them used in statistical teaching and consultancy, the better. We present an example of the usefulness of graphs in studying associations among six soil properties.

Key words: correlation, graphical statistics, information graphics, scatterplot matrix, visualization.

1. Introduction

Statistics is not an easy matter for non-statistical students and researchers. This includes not only the most difficult and complex methods, but also quite regular ones. Populations and samples, distributions, point and interval estimation, hypothesis testing, analysis of variance, regression and so on, are all difficult topics for anyone who is just beginning his or her statistical education. Hence a teacher must do everything that is possible to facilitate the understanding of statistics by students. This applies equally to consultations with non-statistical researchers.

We claim that this might be done by incorporating many more graphical elements into statistics teaching and consultancy than is normally done.

Of course, it is difficult to estimate how much graphics is taught and used, mainly because each teacher has his or her own approach towards statistics and methods of teaching. It is rather unlikely to find two teachers who teach in exactly the same way. This is a good rather than a bad thing, simply because statistics is a kind of art, and no art should be taught in a single uniform way. Nonetheless, graphical elements seem to be far too rare in statistics courses. This can be seen in numerous textbooks on statistics, both international and Polish, where graphs serve only as an insignificant addition. Of course, graphics are an essential element of some statistical methods, including regression and multivariate analysis. However, all too often these methods are reported and – unfortunately – analysed without any use of graphs, which may result in very poor performance of statistical models, and especially in incorrect interpretation.

This paper is a call for the use of more graphics in biological, environmental and agricultural statistics at the elementary level of teaching and consultancy. Thanks to graphs, many statistical issues that are too difficult to understand for non-statisticians can become friendlier, and even more importantly, interpretation and conclusions may become more correct and more conclusive. This is because with graphs one pays more attention to the data rather than just focusing on numerical output from statistical software. We will illustrate these issues with a very simple example of correlation analysis, which is so commonly applied in biological and agricultural literature. If so simple a problem may be misinterpreted, then what is to be said about genuinely difficult ones, those which call for much more sophisticated and complex statistical methods?

This paper, then, aims to discuss how graphs can cure the illness of statistical teaching and consultancy: the lack of communication between a teacher and his or her students, or between a consultant and his or her client.

2. Example: Correlations

Merely focusing on statistical hypothesis testing and forgetting about the data one studies can provide nonsensical results. Consider correlation, for example. Kozak (2008) showed that with a huge sample, a correlation coefficient even smaller than 0.05 can be significant. Does this make any sense? Not much, if one understands this as an indication of the significant linear association between the two variables. Going further, an extremely common way of presenting associations among traits is based on a correlation matrix, which merely reports Pearson's correlation coefficients among all the variables along with their significance indicated by asterisks. (Kozak, 2009 discusses why asterisks should not be used to show significance of correlation and statistical estimates in general.) This standard approach may be very misleading for several reasons. First is the reason mentioned above — testing of correlation may have little sense (see the discussion in Kozak, 2008). Second, the reader is offered no information (or opportunity to obtain it) as to whether there is any nonlinearity among the variables. Third, no information about outliers is offered. Fourth, the reader cannot grasp the whole picture of the associations among the variables. Hence not only can such a correlation matrix provide an unclear picture of the associations among the variables, but the matrix can be incorrectly interpreted.

Table 1 lists Pearson's correlations among six soil traits, taken from the "soil" dataset of the "agricolae" package (de Mendiburu, 2008) of R (R Development Core Team, 2009). The soil traits considered in our analysis are pH, EC (electric conductivity), $CaCO_3$, MO (organic matter), CIC (cation exchange capacity) and P content (the original abbreviations from the dataset are retained). The 13 observations come from different locations. Only two coefficients are significant in Table 1: those between pH and $CaCO_3$ ($P \leq 0.01$), and between P and MO ($P \leq 0.05$); note that the small sample size has quite an impact on this result. One might decide to present only significant correlations, as is sometimes done (e.g. Kobierski, 2004) – see Table 2. Such a table stresses the importance of hypothesis testing, and discards all correlations that may be high although insignificant or close to significant (see the discussion by Kozak, 2008). Table 3, on the other hand, is much better presented than Tables 1 and 2 – instead of asterisks to indicate significance, or providing only significant coefficients, each coefficient is accompanied by the corresponding *p*-value. In this way the reader has more information about the strength of the relationship.

Table 1. Correlation matrix for six soil traits. All correlations are given with the significance indicated with asterisks. Source: "soil" data set, package agricolae of R.

	pH	EC	$CaCO_3$	MO	CIC
EC	0.55				
$CaCO_3$	0.73**	0.32			
MO	–0.33	–0.39	–0.23		
CIC	0.26	0.00	0.30	0.53	
P	0.14	0.46	0.05	0.56*	0.55

*, ** Significant at $p \le 0.05$ and $p \le 0.01$, respectively

Table 2. Correlation matrix for six soil traits. Only significant correlations are given, which in this case gives only two coefficients. Source: "soil" data set, package agricolae of R.

	pH	EC	CaCO3	MO	CIC
EC	ns				
$CaCO_3$	0.73**	ns			
MO	ns	ns	ns		
CIC	ns	ns	ns	Ns	
P	ns	ns	ns	0.56*	ns

*, ** Significant at $p \le 0.05$ and $p \le 0.01$, respectively; ns—nonsignificant

Table 3. Correlation matrix for six soil traits. The corresponding *p*-values are provided in parentheses. Source: "soil" data set, package agricolae of R.

	pH	EC	CaCO3	MO	CIC
EC	0.55 (0.053)				
CaCO3	0.73 (0.005)	0.32 (0.294)			
MO	–0.33 (0.278)	–0.39 (0.187)	–0.23 (0.456)		
CIC	0.26 (0.386)	0.00 (0.988)	0.30 (0.315)	0.53 (0.06)	
P	0.14 (0.651)	0.46 (0.111)	0.05 (0.874)	0.56 (0.045)	0.55 (0.051)

But is the information provided in Table 3 – let alone Tables 1 and 2 – sufficient to get the whole picture of the associations among the variables? It might be, but only under the assumption that only linear relations are possible among the variables within this dataset. Can we make such an assumption? Probably not – we had no prior information that would justify this (even if we did have this information, outliers may occur, sometimes heavily influencing the estimates). Instead, let us draw a set of scatterplots for each pair of variables – this very useful technique is called the scatterplot matrix (Cleveland, 1993, 1994). See Figure 1 for a scatterplot matrix of our six variables; it was constructed with the splom function of the lattice package (Sarkar, 2008) of R (R Development Core Team, 2009). In addition, we have added to each panel a locally weighted regression (*loess*) curve (Cleveland, 1979, 1993, 1994), which aims to show a robust relationship between a row (in terms of a scatterplot matrix) and a column variable; the loess curves were fitted with the re-descending M estimator with Tukey's biweight function. Clearly it would be difficult to claim that all relationships are approximately linear. Of course, besides the intrinsic characteristics of this association, the smallness of the sample and the obvious outliers may have an impact on this problem, but can we simply ignore this fact and choose linear relationships?

See Figure 2. To each panel of the same scatterplot matrix we have added a straight least-square line representing a linear relationship between a row and a column variable. Thus these lines portray the relationships which the correlations in Tables 1 and 3 represent. Clearly, claiming that all these relations are linear would rather be a crude approach to data analysis. Our aim is not to suggest using loess for such types of data (besides, there are other nonparametric regression methods), but rather to recommend careful examination of data using graphical methods before applying statistical analysis.

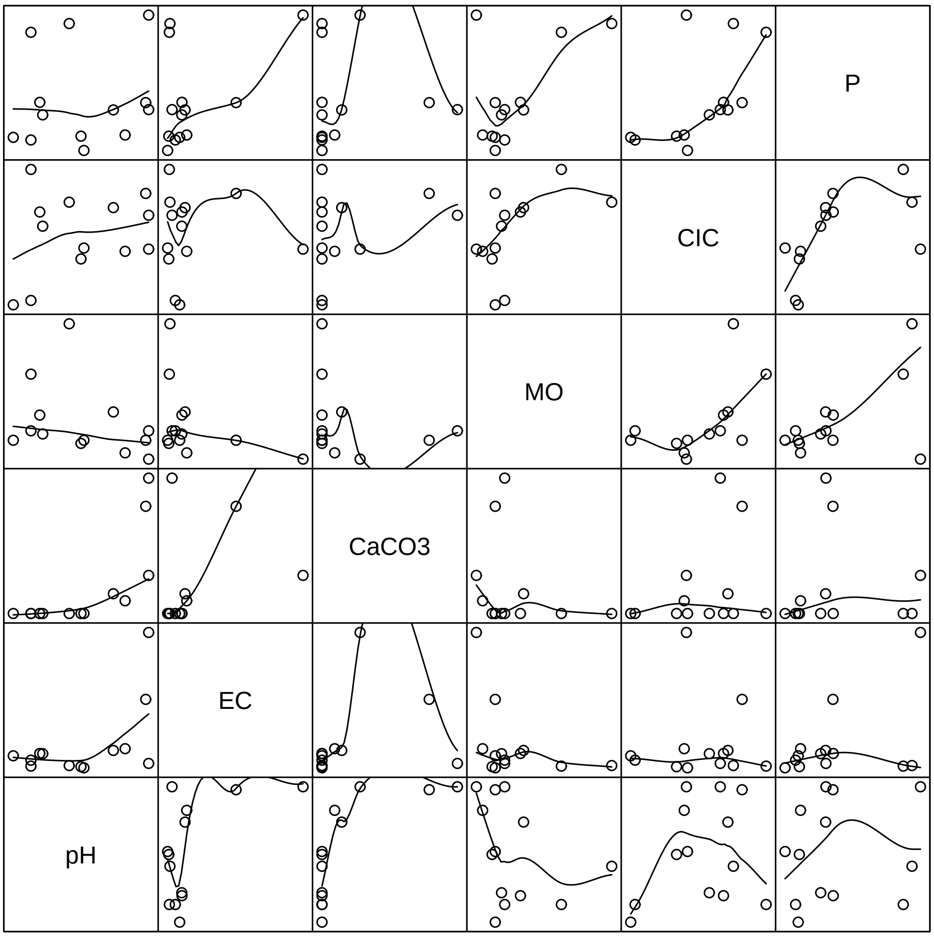

Figure 1. Scatterplot matrix for six soil traits. A locally weighted regression (*loess*) curve has been added to each panel to provide a rough association between the row and column variables. In some panels one can see outliers as well as nonlinearity.

3. Discussion

Graphs have been present in statistics since the very beginning. However, a real milestone in graphical statistics probably came with John W. Tukey's ingenious book on exploratory data analysis (Tukey, 1977). Since then, graphs have been

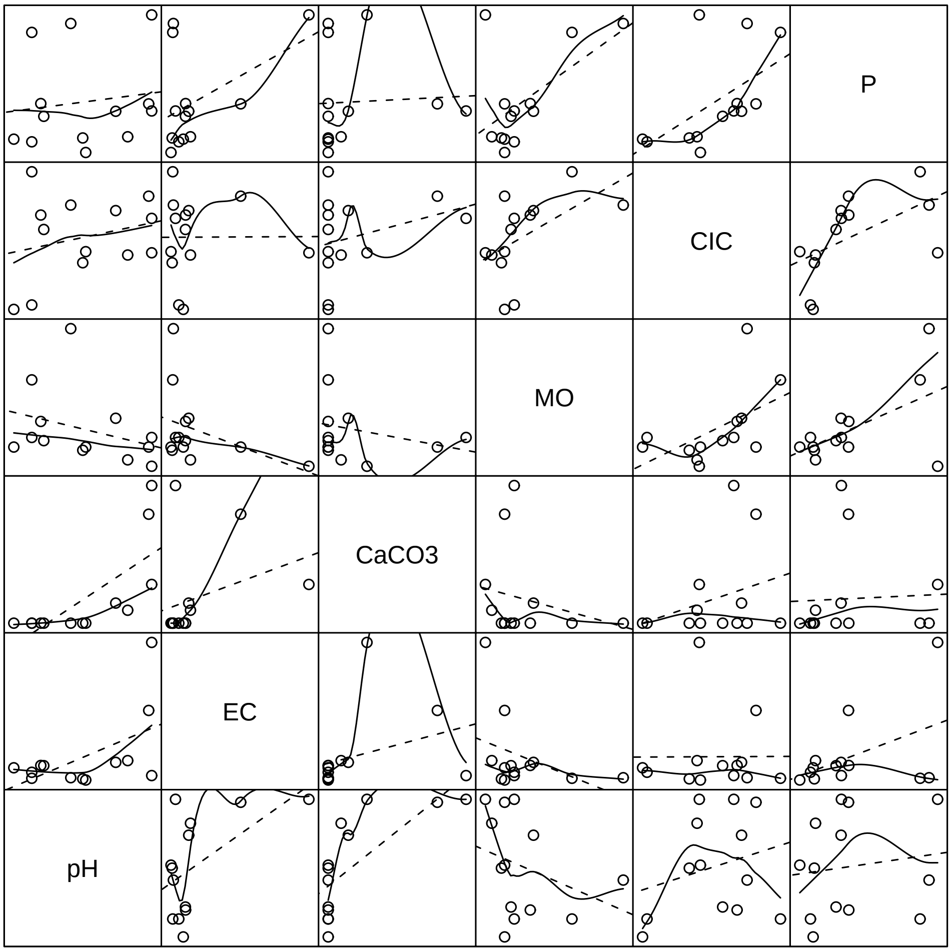

Figure 2. Scatterplot matrix with a locally weighted regression (*loess*) curve from Figure 1, with a superposed dashed least-square line representing a linear relationship between a row and a column variable.

more and more appreciated in statistics and gained more and more interest, including biological, environmental and agricultural applications. Examples include the GGE (Yan, Kang, 2002, Fan et al., 2007) and AMMI (Gauch, 1992) methods for studying genotype-by-environment interactions, which are based on various biplots; various methods of visualization of proteomics and genomics data (e.g., Saldanha, 2004, Brouwer et al., 2009, Carver et al., 2009); PCA (e.g., Nuijten, van Treuren, 2007, Lattoo et al., 2008, Ursem et al., 2008, van Berloo et al., 2008, Xie et al., 2008, Asare et al., 2009, Nicholls, 2009); cluster analysis (e.g., Eisen et al., 1998, Crossa, Franco, 2004); genetic diversity (e.g., Stępień et al., 2007, D'hoop et al., 2008, Lattoo et al., 2008, Yonemori et al., 2008); gene expression (e.g., Mehrian-Shai et al., 2007); multi-trait and multi-environmental QTL analysis (e.g., Malosetti et al., 2008); multidimensional scaling (e.g., van Wezel, Kosters, 2004, Venna, Kaski, 2006, Salmela et al., 2008, Žilinskas, Žilinskas, 2006, Tzeng et al., 2008); and many other multivariate methods (e.g., Debat et al., 2008, Hepperger et al., 2008, Rabelo et al., 2008). Of course, there can be no spatial statistics without graphs (e.g., Ripley, 1981, Grego et al., 2006, Gozdowski et al., 2008, Molin, de Castro, 2008). Other efficient and interesting applications of graphs in various fields of agricultural sciences include Lammel et al. (2007), de Melo et al. (2007), Miele et al. (2007), Bünemann et al. (2008), Ribeiro et al. (2008), Rawlings et al. (2009), and Ribeiro Jr et al. (2009).

The aforementioned articles use some intricate and complex methods and/or graph types for data visualization and analysis. One must nonetheless start with basic methods to learn how graphing works for data analysis. Still, however, some books, even excellent from a statistical point of view, fail to direct readers' attention to graphical approaches. Of course there are other books that stress this very important topic, for example for checking model assumptions and goodness-of-fit (e.g., Quinn and Keough, 2002).

We believe that the above simple example with correlations should convince the reader that data visualization may be a powerful tool for understanding the data and phenomena one wants to study. We also believe that there should be no statistics without visualization, except in rare cases. Of course, graphing is not all roses. One has to spend time on learning useful tools, and the construction of good graphs itself takes time. In addition, Cook and Weisberg (1999) write, "useful graphs must have a *context* induced by associated theory, and… a graph without the well-understood statistical context is hardly worth drawing" (italics original). Of course, Cook and Weisberg have intricate statistical graphics in mind, including residual plots and scatterplot matrices used in the context of multiple regression. Their note is of importance for general data visualization, though — whatever one visualizes, it does have to be set up in the appropriate statistical context, even if this is just exploratory visualization.

Graphing requires some knowledge of relevant software, and not all software is good for this purpose. In this paper, we used R, which is excellent for graphing, but requires quite a bit of time to learn. However, after some time even quite complex graphs become easy to construct. Therefore to what Kozak et al. (2004) wrote about the usefulness of R in biometrical computing, we could add the enormous possibilities it offers in graphical terms.

In this paper we have focused on selected graphs and problems. Visualization, however, offers many more graphical tools to explore data — see for example Cleveland’s books (1993, 1994) to learn about various methods of data visualization, and Tufte (1983, 1990, 1998, 2001, 2006) to learn what can be done with graphs.

Understanding statistics is often very difficult. Graphs can make it easier.

REFERENCES

Asare A.L., Gao Z., Carey V.J., Wang R., Seyfert-Margolis V. (2009): Power enhancement via multivariate outlier testing with gene expression arrays. Bioinformatics 25(1): 48-53.

Brouwer R.W.W., Hijum S.A.F.T., Kuipers O.P. (2009): MINOMICS: visualizing prokaryote transcriptomics and proteomics data in a genomic context. Bioinformatics 25(1): 139-140.

Bünemann E.K., Marschner P., Smernik R.J., Conyers M., McNeill A.M. (2008): Soil organic phosphorus and microbial community composition as affected by 26 years of different management strategies. Biology and Fertility Soils 44: 717-726.

Carver T., Thomson N., Bleasby A., Berriman M., Parkhill J. (2009): DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 25(1), 119-120.

Cleveland W.S. (1979): Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74: 829-836.

Cleveland W.S. (1993): Visualizing data. Summit, NJ: Hobart, USA.

Cleveland W.S. (1994): The elements of graphing data. 2nd ed. Summit, NJ: Hobart, USA.

Cook R.D., Weisberg S. (1999): Graphs in statistical analysis: Is the medium the message? The American Statistician 53(1): 29-37.

Crossa J., Franco J. (2004): Statistical methods for classifying genotypes. Euphytica 137: 19-37.

Debat V., Cornette R., Korol A.B., Nevo E., Soulet D., David J.R. (2008): Multidimensional analysis of *Drosophila* wing variation in Evolution Canyon. Journal of Genetics 87: 407–419.

de Melo S.P., Monteiro F.A., Manfredini D. (2007): Silicate and phosphate combinations for arandu palisadegrass growing on an oxisol. Scientia Agricola (Piracicaba, Braz.) 64(3): 275-281.

de Mendiburu F. (2008): agricolae: Statistical procedures for agricultural research. R package version 1.0-5.

D'hoop B.B., Paulo M.J., Mank R.A., van Eck H.J., van Eeuwijk F.A. (2008): Association mapping of quality traits in potato (*Solanum tuberosum* L.). Euphytica 161: 47-60.

Eisen M.B., Spellman P.T., Brown P.O., Botstein D. (1998): Cluster analysis and display of genome-wise expression patterns. Proceedings of the National Academy of Sciences USA 95: 14863-14868.

Fan X.-M., Kang M. S., Chen H., Zhang Y., Tan J., Xu C. (2007): Yield stability of maize hybrids evaluated in multi-environment trials in Yunnan, China. Agronomy Journal 99: 220-228.

Gauch H.G. (1992): Statistical analysis of regional yield trials: AMMI analysis of factorial designs. Elsevier, Amsterdam.

Gozdowski D., Roszkowska-Mądra B., Mądry W. (2008): Crop diversity at the gmina level and its causes in the Podlasie district of Poland. Communications in Biometry and Crop Science 3(2): 72-79.

Grego C.R., Viera S.R., Antonio A.M., Rosa S.C.D. (2006): Geostatistical analysis for soil moisture content under the no tillage cropping system. Scientia Agricola (Piracicaba, Braz.) 63(4): 341-350.

Hepperger C., Mannes A., Merz J., Peters J., Dietzel S. (2008): Three-dimensional positioning of genes in mouse cell nuclei. Chromosoma 117: 535-551.

Kobierski M. (2004): Zawartość miedzi, cynku, manganu i żelaza w glebach sadów jabłoniowych w 27. i 30. roku ich użytkowania. Acta Scientiarum Polonorum, Hortorum Cultus 3(2): 161-170.

Kozak M. (2009): Asterisks — friends or foes of statistics? *Teaching Statistics* (in press, after proof).

Kozak M. (2008): Correlation coefficient and the fallacy of statistical hypothesis testing. Current Science 95(9): 1121-1122.

Kozak M., Laudański Z., Zieliński A. (2004): Biometrical computing with R. Colloquium Biometryczne 34A: 51-64.

Lammel D.R., Brancalion P.H.S., Dias C.T.S., Cardoso E.J.B.N. (2007): Rhizobia and other legume nodule bacteria richness in *Brazilian Araucaria angustifolia forest.* Scientia Agricola (Piracicaba, Braz.) 64(4)*:* 400-408.

Lattoo S.K., Dhar R.S., Khan S., Bamotra S., Bhan M.K., Dhar A.K., Gupta K.K. (2008): Comparative analysis of genetic diversity using molecular and morphometric markers in *Andrographis paniculata* (Burm. f.) Nees. Genetic Resources and Crop Evolution 55: 33-43.

Malosetti M., Ribaut J.M., Vargas M., Crossa J., van Eeuwijk F.A. (2008): A multi-trait multi-environment QTL mixed model with a application to drought and nitrogen stress trails in maize (*Zea mays* L.). Euphytica 161: 241-257.

Mehrian-Shai R., Chen C.D., Shi T., Horvath S., Nelson S.F., Reichardt J.K.V., Sawyers C.L. (2007): Insulin growth factor-binding protein 2 is a candidate biomarker for PTEN status and PI3K/Akt pathway activation in glioblastoma and prostate cancer. Proceedings of the National Academy of Sciences USA 104: 5563-5568.

Miele M., Coldebella A., Waquil P.D., Miele A. (2007): Segments of competition in South Brazilian wineries. Scientia Agricola (Piracicaba, Braz.) 64(3): 227-234.

Molin J.P., de Castro C.N. (2008): Establishing management zones using soil electrical conductivity and other soil properties by the fuzzy clustering technique. Scientia Agricola (Piracicaba, Braz.) 65(6): 567-573.

Nicholls K.H. (2009): A multivariate statistical evaluation of the "acolla-complex" of Corythionella species, including a description of *C. darwini* n. sp. (Rhizopoda: Filosea or Rhizaria: Cercozoa). European Journal of Protistology 45(3): 183-192.

Nuijten E., van Treuren R. (2007): Spatial and temporal dynamics in genetic diversity in upland rice and late millet (*Pennisetum glaucum* (L.) R. Br.) in The Gambia. Genetic Resources and Crop Evolution 54: 989-1009.

Quinn G.P., Keough M.J. (2002): Experimental design and data analysis for biologists. Cambridge University Press, Cambridge.

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.

Rabelo S.C., Filho R.M., Costa A.C. (2008): A comparison between lime and alkaline hydrogen peroxide pretreatments of sugarcane bagasse for ethanol production. Applied Biochemistry and Biotechnology 144: 87-100.

Rawlings B.G., Webster R., Tye A.M., Lawley R., O'Hara S.L. (2009): Estimating particle-size fractions of soil dominated by silicate minerals from geochemistry. European Journal of Soil Science 60: 116-126.

Ribeiro R.V., de Souza Rolim G., de Azevedo F.A., Machado E.C. (2008): 'Valancia' sweet orange tree flowering evaluation under field conditions. Scientia Agricola (Piracicaba, Braz.) 65(4): 389-396.

Ribeiro Jr P.J., Viola D.N., Demétrio C.G.B., Manly B.F., Fernandes O.D. (2009): Spatial pattern detection modeling of thrips (*Thrips tabaci*) on onion fields. Scientia Agricola (Piracicaba, Braz.) 66(1): 90-99.

Ripley B.D. (1981): Spatial statistics. Wiley, New York.

Saldanha A.J. (2004): Java Treeview-extensible visualization of microarray data. Bioinformatics 20(17): 3246-3248.

Salmela E., Lappalainen T., Fransson I., Andersen P.M., Dahlman-Wright K., Fiebig A., Sistonen P., Savontaus M.-L., Schreiber S., Kere J., Lahermo P. (2008): Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS ONE* 3(10), e3519. doi:10.1371/journal.pone.0003519

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

Stępień Ł., Mohler V., Bocianowski J., Koczyk G. (2007): Assessing genetics diversity of Polish wheat (*Triticum aestivum*) varieties using microsatellite markers. Genetic Resources and Crop Evolution 54: 1499-1506.

Tufte E.R. (1983, 2001): The visual display of quantitative information. 1st and 2nd ed. Graphics Press LLC, Cheshire.

Tufte E.R. (1990): Envisioning information. Graphics Press LLC, Cheshire.

Tufte E.R. (1998): Visual explanations: Images and quantities, evidence and narrative. Graphics Press LLC, Cheshire.

Tufte E.R. (2006): Beautiful evidence. Graphics Press LLC, Cheshire.

Tukey J.W. (1977): Exploratory data analysis. Addison-Wesley, Reading, Mass.

Tzeng J., Lu H.H.-S., Li W.-H. (2008): Multidimensional scaling for large genomic data sets. BMC Bioinformatics 9, 179. doi:10.1186/1471-2105-9-179.

Ursem R., Tikunov Y., Bovy A., van Berloo R., van Eeuwijk F.A. (2008): A correlation network approach to metabolic data analysis for tomato fruits. Euphytica 161: 181-193.

van Berloo R., van Heusden S., Bovy A., Meijer-Dekens F., Lindhout P., van Eeuwijk F.A. (2008): Genetic research in a public-private research consortium: prospects for indirect use of Elite breeding germplasm in academic research. Euphytica 161, 293-300.

van Wezel M.C., Kosters W.A. (2004): Nonmetric multidimensional scaling: Neural networks versus traditional techniques. Intelligent Data Analysis 8: 601-613.

Venna J., Kaski S. (2006): Local multidimensional scaling. Neural Networks 19: 889-899.

Xie L.-J., Ye X.-Q., Liu D.-H., Ying Y.-B. (2008): Application of principal component-radial basis function neural networks (PC-RBFNN) for the detection of water-adulterated bayberry juice by near-infrared spectroscopy. Journal of Zhejiang University Science B 9(12): 982-989.

Yan W., Kang M.S. (2002): GGE Biplot analysis: A graphical tool for breeders, Geneticists, and Agronomists. CRC Press, Boca Raton, FL.

Yonemori K., Honsho C., Kitajima A., Aradhya M., Giordani E., Bellini E., Parfitt D.E. (2008): Relationship of European persimmon (*Diospyros kaki* Thunb.) cultivars to Asian cultivars, characterized using AFLPs. Genetic Resources and Crop Evolution 55: 81-89.

Žilinskas A., Žilinskas J. (2006): On multidimensional scaling with Euclidean and city block metrics. Technological and Economic Development of Economy XII (1): 69-75.