Using R to Perform the AMMI Analysis on Agriculture Variety Trials

by Andrea Onofri & Egidio Ciriciofolo

Introduction

Field agricultural experiments are generally planned to evaluate the actual effect produced by man-made chemical substances or human activities on crop yield and quality, environmental health, farmers' income and so on. Field experiments include the testing of new and traditional varieties (genotypes), fertilizers (types and doses), pesticides (types and doses) and cultural practices. With respect to greenhouse or laboratory experiments, field trials are much more strongly subjected to environmental variability that forces researchers into repeating experiments across seasons and/or locations. A significant 'treatment x environment' interaction may introduce some difficulties in exploring the dataset, summarizing results and determining which treatment (genotype, herbicide, pesticide...) was the best.

In such conditions, the AMMI (Additive Main effect Multiplicative Interaction) analysis has been proposed as an aid to visualize the dataset and explore graphically its pattern and structure (Gollob, 1968; Zobel et al., 1988); this technique has received a particular attention from plant breeders (see for example Abamu and Alluri, 1998; Annichiarico et al., 1995; Annichiarico, 1997; Ariyo, 1998) and recently it has been stated as superior to other similar techniques, such as the GGE (Gauch, 2006). Unfortunately, such technique is not yet very well exploited by agricultural scientists, who often prefer a more traditional approach to data analysis, based on ANOVA and multiple comparison testing. Without disregarding the importance of such an approach, one cannot deny that sometimes this does not help unveil the underlying structure of experimental data, which may be more important than hypothesis testing, especially at the beginning of data analyses (Exploratory Data Analysis; NIST/SEMATECH, 2004) or at the very end, when graphs have to be drawn for publication purposes.

To make more widespread the acceptance and the use of such powerful tool within agronomists, it is necessary to increase the availability of both practical information on how to perform and interpret an AMMI analysis and simple software tools that give an easily understandable output, aimed at people with no specific and deep statistical training, such as students and field technicians.

The aim of this paper was to show how R can be easily used to perform an AMMI analysis and produce 'biplots', as well as to show how these tools can be very useful within variety trials in agriculture.

Some basic statistical aspects

The AMMI analysis combines the ANalysis OF VAriance (ANOVA) and the Singular Value Decomposition (SVD) and it has been explained in detail by Gollob (1968). If we specifically refer to a variety trial, aimed at comparing the yield of several genotypes in several environments (years and/or locations), the ANOVA partitions the total sum of squares into two main effects (genotypes and environments) plus the interaction effect (genotypes x environments). This latter effect may be obtained by taking the observed averages for each 'genotype x environment' combination and doubly-centering them (i.e., subtracting to each data the appropriate genotype and environment means and adding back the grand mean). The interaction effect is arranged on a two-way matrix γ (one row for each genotype and one column for each environment) and submitted to SVD, as follows:

$$\gamma = \sum_{i=1}^{r} \lambda_i \cdot g_{ik} \cdot e_{ij} \qquad (1)$$

where r is the rank of γ, λ_i is the singular value for principal component i, g_{ik} is the eigenvector score for genotype k and Principal Component (PC) i (left singular vector), while e_{ij} is the eigenvector score for environment j and PC i (right singular vector). If PC scores are multiplied by the square root of the singular value, equation 1 is transformed into:

$$\gamma = \sum_{i=1}^{r} \left(\lambda_i^{0.5} \cdot g_{ik}\right)\left(\lambda_i^{0.5} \cdot e_{ij}\right) = \sum_{i=1}^{r} G_{ik} \cdot E_{ij} \qquad (2)$$

In this way the additive interaction in the ANOVA model is obtained by multiplication of genotype PC scores by environment PC scores, appropriately scaled. If a reduced number of PCs is used ($r = 1$ or 2, typically) a dimensionality reduction is achieved with just a small loss in the descriptive ability of the model. This makes it possible to plot the interaction effect, via the PC scores for genotypes and environments. Such graphs are called biplots, as they contain two kinds of data; typically, a AMMI1 and a AMMI2 biplots are used: the AMMI1 biplot has main effects (average yields for genotypes and environments) on the x-axis and PC1 scores on the y-axis, while the AMMI2 biplot has PC1 scores on the x-axis and PC2 scores on the y-axis.

 ISSN 1609-3631

Table 1: Field averages (three replicates) for six genotypes compared in seven years.

Genotype	1996	1997	1998	1999	2000	2001	2002	Average
COLOSSEO	6.35	6.46	6.70	6.98	6.44	7.07	4.90	6.41
CRESO	5.60	6.09	6.13	7.13	6.08	6.45	4.33	5.97
DUILIO	5.64	8.06	7.15	7.99	5.18	7.88	4.24	6.59
GRAZIA	6.26	6.74	6.35	6.84	4.75	7.30	4.34	6.08
IRIDE	6.04	7.72	6.39	7.99	6.05	7.71	4.96	6.70
SANCARLO	5.70	6.77	6.81	7.41	5.66	6.67	4.50	6.22
SIMETO	5.08	7.19	6.44	7.07	4.82	7.55	3.34	5.93
SOLEX	6.14	6.39	6.44	6.87	5.45	7.52	4.79	6.23
Average	5.85	6.93	6.55	7.29	5.55	7.27	4.42	6.27

The dataset

To show how the AMMI analysis can be easily performed with R, we will use a dataset obtained from a seven-years field trial on durum wheat, carried out from 1996 to 2003 in central Italy, on a randomised block design with three replicates. For the present analysis, eight genotypes were chosen, as they were constantly present throughout the years (Colosseo, Creso, Duilio, Grazia, Iride, Sancarlo, Simeto, Solex). Yield data referred to the standard humidity content of 13% (Tab. 1) have been previously published in Belocchi et al. (2003), Ciriciofolo et al. (2002); Ciriciofolo et al. (2001); Desiderio et al. (2000), Desiderio et al. (1999), Desiderio et al. (1998), Desiderio et al. (1997). The interaction matrix (which is submitted to SVD) is given in table 2.

The AMMI with R

To perform the AMMI analysis, an R function was defined, as shown on page 17.

The `AMMI()` function requires as inputs a vector of genotype codes (factor), a vector of environment codes (factor), a vector of block codes (factor) and a vector of yields (numerical). PC is the number of PCs to be considered (set to 2 by default) and biplot is the type of biplot to be drown (1 for AMM1 and 2 for AMMI2). It should be noted that the script is very elementary and that it does not use any difficult function or programming construct. It was simply coded by translating the algebraic procedure proposed by Gollob (1968) into R statements, which is a very easy task, even without a specific programming training. Wherever possible, built-in R functions were used, to simplify the coding process and to facilitate the adaptation of the script to other kinds of AMMI models.

The first part uses the function `tapply()` to calculate some descriptive statistics, such as genotype means, environment means and 'genotype x environment' means, which are all included in the final output.

The second part uses the function `aov()` to perform the ANOVA by using a randomised block design repeated in different environments with a different randomisation in each environment (LeClerg et al., 1962). The interaction matrix γ is calculated by using the function `model.tables()` applied to the output of the function `aov()`; the way the R script is coded, the interaction matrix is actually the transpose of the matrix shown in table 2, but this does not change much in terms of the results. The interaction matrix is then submitted to SVD, by using the built-in R function `svd()`.

The significant PCs are assessed by a series of F tests as shown by Zobel et al. (1988) and PC scores, genotype means and environment means are used to produce biplots, by way of the functions `plot()` and `points()`.

Table 2: Interaction effects for the dataset in table 1.

Genotype	1996	1997	1998	1999	2000	2001	2002
COLOSSEO	0.35	-0.62	0.00	-0.45	0.74	-0.34	0.33
CRESO	0.04	-0.54	-0.13	0.14	0.82	-0.52	0.20
DUILIO	-0.54	0.80	0.28	0.38	-0.70	0.29	-0.51
GRAZIA	0.60	-0.01	-0.02	-0.27	-0.62	0.21	0.10
IRIDE	-0.24	0.37	-0.59	0.28	0.07	0.01	0.11
SANCARLO	-0.10	-0.11	0.31	0.17	0.16	-0.55	0.12
SIMETO	-0.43	0.60	0.23	0.13	-0.40	0.62	-0.75
SOLEX	0.33	-0.50	-0.07	-0.38	-0.07	0.29	0.40

```
$Additive_ANOVA
                         Df  Sum Sq Mean Sq  F value    Pr(>F)
Environments              6 159.279  26.547 178.3996 < 2.2e-16 ***
Genotypes                 7  11.544   1.649  11.0824 2.978e-10 ***
Blocks(Environments)     14   3.922   0.280   1.8826   0.03738 *
Environments x Genotypes 42  27.713   0.660   4.4342 6.779e-10 ***
Residuals                98  14.583   0.149
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Mult_Interaction
     Effect        SS DF        MS         F        Prob.
1       PC1 18.398624 12 1.5332187 10.303612 7.958058e-13
2       PC2  5.475627 10 0.5475627  3.679758 3.339881e-04
3       PC3  1.961049  8 0.2451311  1.647342 1.212529e-01
4 Residuals  1.877427 12 0.1564522  1.051398 4.094193e-01

$Environment_scores
            PC1         PC2         PC3
1996  0.4685599 -0.62599974  0.01665148
1997 -0.8859669  0.21085535 -0.19553672
1998 -0.1572887 -0.00567589  0.80162642
1999 -0.3139136  0.51881710 -0.13286326
2000  0.8229290  0.59868592 -0.03330554
2001 -0.5456613 -0.49726356 -0.18138908
2002  0.6113417 -0.19941917 -0.27518331

$Genotype_scores
                PC1         PC2           PC3
COLOSSEO  0.74335025 -0.02451524  0.1651197989
CRESO     0.63115567  0.47768803 -0.0001969871
DUILIO   -0.87632103  0.17923645  0.1445152042
GRAZIA   -0.07625519 -0.74659598 -0.0108977060
IRIDE    -0.12683903  0.28634343 -0.7627600696
SANCARLO  0.18186612  0.35076556  0.3753706117
SIMETO   -0.78109997  0.04751457  0.1740113396
SOLEX     0.30414317 -0.57043681 -0.0851621918
```

Figure 1: Results from ANOVA and AMMI analyses.

Results

Results (Fig. 1) show a highly significant 'genotypes x environments' interaction (GE) on the ANOVA, that does not permit to define an overall ranking of varieties across environments.

The SVD decomposition of the interaction matrix was performed by extracting three PCs, though only the first two are significant. It is possible to observe that the first PC accounts for 66% of the interaction sum of squares, while the second one accounts for an additional 20%.

The AMMI1 biplot shows contemporarily main effects (genotypes and environments average yields) and interaction, as PC1 scores (Fig. 2). This graph is relevant as it accounts for 87% of total data variability.

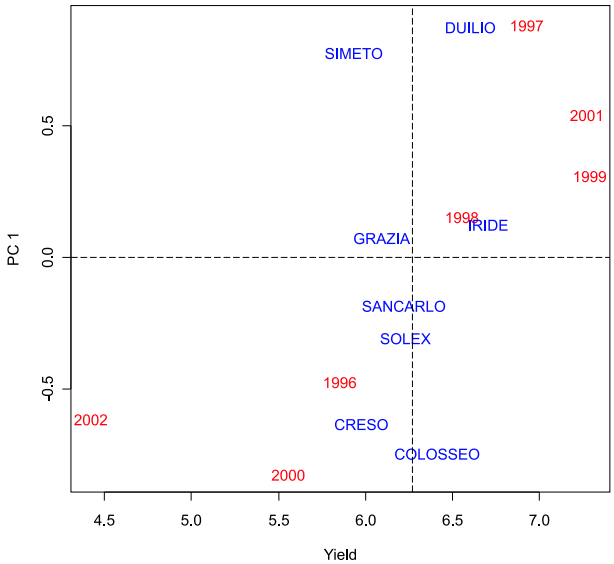

Figure 2: AMMI1 biplot.

```
AMMI <- function(variety, envir, block, yield, PC = 2, biplot = 1) {
## 1 - Descriptive statistics
  overall.mean <- mean(yield)
  envir.mean <- tapply(yield, envir, mean)
  var.mean <- tapply(yield, variety, mean)
  int.mean <- tapply(yield, list(variety,envir), mean)
  envir.num <- length(envir.mean)
  var.num <- length(var.mean)
## 2 - ANOVA (additive model)
  variety <- factor(variety)
  envir <- factor(envir)
  block <- factor(block)
  add.anova <- aov(yield ~ envir + block %in% envir + variety + envir:variety)
  modelTables <- model.tables(add.anova, type = "effects", cterms = "envir:variety")
  int.eff <- modelTables$tables$"envir:variety"
  add.anova.residual.SS <- deviance(add.anova)
  add.anova.residual.DF <- add.anova$df.residual
  add.anova.residual.MS <- add.anova.residual.SS/add.anova.residual.DF
  anova.table <- summary(add.anova)
  row.names(anova.table[[1]]) <- c("Environments", "Genotypes", "Blocks(Environments)",
                                   "Environments x Genotypes", "Residuals")
## 3 - SVD
  dec <- svd(int.eff, nu = PC, nv = PC)
  if (PC > 1) {
    D <- diag(dec$d[1:PC])
  } else {
    D <- dec$d[1:PC]
  }
  E <- dec$u %*% sqrt(D)
  G <- dec$v %*% sqrt(D)
  Ecolnumb <- c(1:PC)
  Ecolnames <- paste("PC", Ecolnumb, sep = "")
  dimnames(E) <- list(levels(envir), Ecolnames)
  dimnames(G) <- list(levels(variety), Ecolnames)
## 4 - Significance of PCs
  numblock <- length(levels(block))
  int.SS <- (t(as.vector(int.eff)) %*% as.vector(int.eff))*numblock
  PC.SS <- (dec$d[1:PC]^2)*numblock
  PC.DF <- var.num + envir.num - 1 - 2*Ecolnumb
  residual.SS <- int.SS - sum(PC.SS)
  residual.DF <- ((var.num - 1)*(envir.num - 1)) - sum(PC.DF)
  PC.SS[PC + 1] <- residual.SS
  PC.DF[PC + 1] <- residual.DF
  MS <- PC.SS/PC.DF
  F <- MS/add.anova.residual.MS
  probab <- pf(F, PC.DF, add.anova.residual.DF, lower.tail = FALSE)
  percSS <- PC.SS/int.SS
  rowlab <- c(Ecolnames, "Residuals")
  mult.anova <- data.frame(Effect = rowlab, SS = PC.SS, DF = PC.DF, MS = MS, F = F, Prob. = probab)
## 5 - Biplots
  if (biplot == 1) {
    plot(1, type = 'n', xlim = range(c(envir.mean, var.mean)), ylim = range(c(E[,1], G[,1])), xlab = "Yield",
     ylab = "PC 1")
    points(envir.mean, E[,1], col = "red", lwd = 5)
    plot(1, type = 'n', xlim = range(c(envir.mean, var.mean)), ylim = range(c(E[,1], G[,1])), xlab = "Yield",
      ylab = "PC 1")
    points(envir.mean, E[,1], "n", col = "red", lwd = 5)
    text(envir.mean, E[,1], labels = row.names(envir.mean), adj = c(0.5, 0.5), col = "red")
    points(var.mean, G[,1], "n", col = "blue", lwd = 5)
    text(var.mean, G[,1], labels = row.names(var.mean), adj = c(0.5, 0.5), col = "blue")
    abline(h = 0, v = overall.mean, lty = 5)
  } else {
    plot(1, type = 'n', xlim = range(c(E[,1], G[,1])), ylim = range(c(E[,2], G[,2])), xlab = "PC 1",
     ylab = "PC 2")
    points(E[,1], E[,2], "n",col = "red", lwd = 5)
    text(E[,1], E[,2], labels = row.names(E),adj = c(0.5,0.5),col = "red")
    points(G[,1],G[,2], "n", col = "blue", lwd = 5)
    text(G[,1], G[,2], labels = row.names(G),adj = c(0.5, 0.5), col = "blue")
  }
## 6 - Other results
  list(Genotype_means = var.mean, Environment_means = envir.mean,  Interaction_means = int.mean,
       Additive_ANOVA = anova.table, Mult_Interaction = mult.anova, Environment_scores = E,
       Genotype_scores = G)
}
```

To read this biplot, it is necessary to remember that genotypes and environments on the right side of the graph shows yield levels above the average. Besides, genotypes and environments laying close to the x-axis (PC 1 score close to 0) did not interact with each other, while data with positive/negative score on y-axis interacted positively with environments characterised by a score of same sign.

Indeed, environmental variability was much higher than genotype variability (Fig. 2, see also the ANOVA in Fig. 1). Iride showed the highest average yield and did not interact much with the environment (PC1 score close to 0). Duilo ranked overall second, but showed a high interaction with the environment, i.e., its yield was above the average in 1997 (first ranking), 2001 (first ranking) and 1999 (second ranking), while it was below the average in 1996, 2000 and 2002. Colosseo gave also a good average yield, but its performances were very positive in 1996, 2000 and 2002, while they were below the average in 1997, 2000 and 2002.

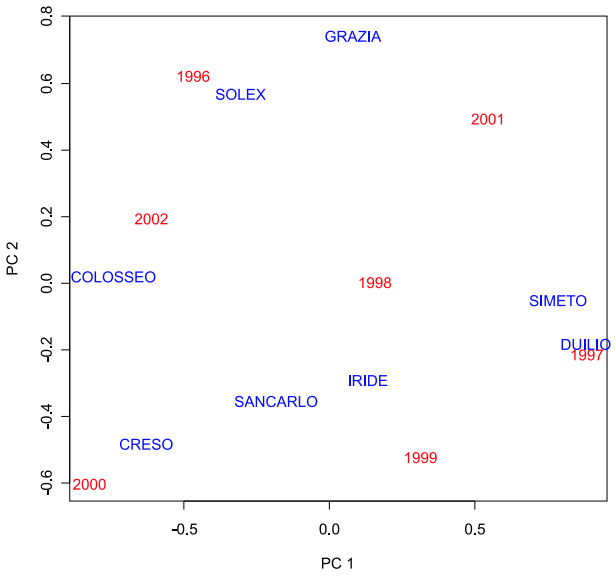

Figure 3: AMMI2 biplot.

The AMMI2 biplot (Fig. 3) is more informative on the GE interaction as it accounts for 86% of the sum of squares of this latter effect. Remember that genotypes and environments in the center of the graph did not show a relevant interaction, while genotypes and environment lying close on the external parts of the graph interacted positively. Duilio and Simeto were particularly brilliant in 1997 (compared to their average performances; notice in tab. 1 that Simeto was the third in this year, which is very good compared to its seventh position on the ranking based on average yield). Solex and Grazia were brilliant in 1997 (they were third and second respectively, in spite of the eighth and fifth ranking based on average yield). Likewise, Creso and Colosseo were the best in 2000 and 2002, while Iride and Sancarlo interacted positively with 1999.

Discussion and conclusions

The above example should be discussed with reference to two main aspects: the AMMI analysis and the use of R. Concerning the first aspect, the example confirms the graphical power of the biplots; indeed, all the above comments are just an excerpt of what can be easily grasped at first sight from the AMMI1 and AMMI2 biplots. It is worth to notice that obtaining such information from table 1 is not as immediate and quick. Of course, the AMMI analysis should be followed by other procedures to explore the relationship between the behaviour of each variety and the environmental conditions of each year.

It is also necessary to mention that in the present example the analyses were aimed only at graphically exploring the underlying structure of the dataset. In other cases, whenever hypothesis testing is more important, the F procedure employed on the script may be too liberal and other techniques may be better suited to evaluate the significance of PCs (Cornelius, 1993).

Concerning R, the above example confirms that this environment can be easily used to perform statistical analyses in the agricultural field. Thanks to its ability to deal with linear models and to the facilities for matrix manipulation, it is very easy to accomplish also rather complex statistical tasks, such as the AMMI analysis. Indeed, calculations can be performed having in mind the usual algebraic notation, as one can find in statistical literature, without a deep knowledge of programming constructs. Indeed, this script has been coded in an elementary fashion, following the calculation pattern proposed by Gollob (1968) and including some built-in R functions when possible.

Of course, it is necessary to mention that this elementary coding style may be useful for simple scripts, but should not be regarded as optimal, especially for more advanced applications. Indeed, in such cases an object-oriented approach is much more advisable to exploit the statistical power of R. In any case, elementary scripts such this one may be always used as the starting point to perform other types of statistical analyses. In particular, with slight modifications, this script (available on: `www.unipg.it/~onofri/software.htm`) could be used to draw the GGE biplot, that has received a certain attention in the last years (Yan and Tinker, 2005).

However, when re-using this script, one should bear in mind some limitations. Indeed, it is important to notice that this script has been aimed at a specific experimental design (completely randomised block experiment repeated across years or environments), as commonly found in field variety trials. Other designs will require some adaptations into the code and unbalanced designs (especially those with missing combinations of genotypes and environments) should not be analysed with this script.

Furthermore, the 'environment' effect has been considered as 'fixed' and changes to the code should be made in case it should be considered as 'random'.

In spite of the the above limitations, it is clear that also users with a limited background in computer programming (which is often the case in agriculture) can benefit from the use of R: an elementary knowledge of R statements and functions is already enough to perform also the 'less traditional' statistical analysis, with a very slight effort.

Bibliography

F. J. Abamu and K. Alluri. Ammi analysis of rainfed lowland rice (oryza sativa) trials in nigeria. *Plant Breeding*, 117(4):395–397, 1998. 14

P. Annichiarico. Additive main effects and multiplicative interaction (ammi) analysis of genotype-location interaction in variety trials repeated over years. *Theorethycal applied genetics*, 94:1072–1077, 1997. 14

P. Annichiarico, F. Bozzo, G. Parente, F. Gusmeroli, V. Mair, O. Marguerettaz, and D. Orlandi. Analysis of adaptation of grass/legume mixtures to italian alpine and subalpine zones through an additive main effects and multiplicative interaction model. *Grass and Forage Science*, 50:405–413, 1995. 14

O. J. Ariyo. Use of additive main effects and multiplicative interaction model to analyse multilocation soybean varietal trials. *J.Genet. Breed*, (53):129–134, 1998. 14

A. Belocchi, M. Fornara, E. Ciriciofolo, E. Biancolatte, S. D. Puglia, G. Olimpieri, V. Vecchiarelli, E. Gosparini, C. Piccioni, and E. Desiderio. Scelta delle varietà di frumento duro: risultati delle prove varietali 2002-2003 in Lazio ed Umbria. *L'Informatore Agrario, supplemento 1*, 2003. 15

E. Ciriciofolo, A. Belocchi, E. Biancolatte, S. D. Puglia, M. Fornara, G. Olimpieri, C. Piccioni, and E. Desiderio. Scelta delle varietà di frumento duro: risultati delle prove varietali 2000-2001 in Lazio ed Umbria. *L'Informatore Agrario, supplemento 1*, 2001. 15

E. Ciriciofolo, A. Belocchi, E. Biancolatte, S. D. Puglia, M. Fornara, G. Olimpieri, C. Piccioni, and E. Desiderio. Scelta delle varietà di frumento duro: risultati delle prove varietali 2001-2002 in Lazio ed Umbria. *L'Informatore Agrario, supplemento 1*, 2002. 15

P. L. Cornelius. Statistical tests and retention of terms in the Additive Main Effects and Multiplicative interaction model for cultivar trials. *Crop Science*, 33:1186–1193, 1993. 18

E. Desiderio, A. Belocchi, E. Biancolatte, E. Ciriciofolo, S. D. Puglia, G. Olimpieri, C. Piccioni, and M. Fornara. Scelta delle varietà di frumento duro: risultati delle prove varietali 1996-1997 in Lazio ed Umbria. *L'Informatore Agrario, supplemento 1*, 1997. 15

E. Desiderio, E. Ciriciofolo, A. Belocchi, E. Biancolatte, S. D. Puglia, G. Olimpieri, C. Piccioni, M. Fornara, and M. Magini. Scelta delle varietà di frumento duro: risultati delle prove varietali 1997-1998 in Lazio ed Umbria. *L'Informatore Agrario, supplemento 1*, 1998. 15

E. Desiderio, E. Ciriciofolo, A. Belocchi, E. Biancolatte, S. D. Puglia, G. Olimpieri, C. Piccioni, M. Fornara, and M. Magini. Scelta delle varietà di frumento duro: risultati delle prove varietali 1998-1999 in Lazio ed Umbria. *L'Informatore Agrario, supplemento 1*, 1999. 15

E. Desiderio, E. Biancolatte, E. Ciriciofolo, S. D. Puglia, M. Fornara, G. Olimpieri, C. Piccioni, and L. Frongia. Scelta delle varietà di frumento duro: risultati delle prove varietali 1999-2000 in Lazio ed Umbria. *L'Informatore Agrario, supplemento 1*, 35: 52–55, 2000. 15

H. F. Gollob. A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika*, 33(1):73–114, 1968. 14, 15, 18

H. G. Gauch. Statistical analysis of yield trials by AMMI and GGE. *Crop Science*, 46(4):1488–1500, 2006. 14

E. L. LeClerg, W. H. Leonard, and A. G. Clark. *Field Plot Technique*. Burgess Publishing Company, Minneapolis, Minnesota, 1962. 15

NIST/SEMATECH. *E-Handbook of Statistical Methods. 1: Exploratory Data Analysis*. `http://www.itl.nist.gov/div898/handbook/`, 2004. 14

W. Yan and N. A. Tinker. An integrated biplot analysis system for displaying, interpreting, and exploring genotype × environment interaction. *Crop Science*, 45(3):1004–1016, 2005. 18

R. W. Zobel, M. J. Wright, and H. G. Gauch. Statistical analysis of a yield trial. *Agronomy Journal*, 80: 388–393, 1988. 14, 15

Andrea Onofri & Egidio Ciriciofolo
Department of Agricultural and Environmental Science
University of Perugia
`onofri@unipg.it`
`egicir@unipg.it`