Agricolae - Ten years of an Open source Statistical tool for experiments in Breeding, agriculture and biology

Felipe Mendiburu, Reinhard Simon

Plant breeders and educators working with the International Potato Center (CIP) needed freely available statistical tools. In response, we created first a set of scripts for specific tasks using the open source statistical software R. Based on this we eventually compiled the R package agricolae as it covered a niche. Here we describe for the first time its main functions in the form of an article. We also review its reception using download statistics, citation data, and feedback from a user survey. We highlight usage in our extended network of collaborators. The package has found applications beyond agriculture in fields like aquaculture, ecology, biodiversity, conservation biology and cancer research. In summary, the package agricolae is a well established statistical toolbox based on R with a broad range of applications in design and analyses of experiments also in the wider biological community .

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1404v1 | CC-BY 4.0 Open Access | rec: 29 Sep 2015, publ: 29 Sep 2015

Agricolae - ten years of an open source statistical tool for experiments in breeding, agriculture and biology

Felipe de Mendiburu[1, 2] **and Reinhard Simon**[1]

[1]**International Potato Center (CIP)**
[2]**Universidad Agraria La Molina (UNALM)**

ABSTRACT

Plant breeders and educators working with the International Potato Center (CIP) needed freely available statistical tools. In response, we created first a set of scripts for specific tasks using the open source statistical software R. Based on this we eventually compiled the *R* package *agricolae* as it covered a niche. Here we describe for the first time its main functions in the form of an article. We also review its reception using download statistics, citation data, and feedback from a user survey. We highlight usage in our extended network of collaborators. The package has found applications beyond agriculture in fields like aquaculture, ecology, biodiversity, conservation biology and cancer research. In summary, the package *agricolae* is a well established statistical toolbox based on R with a broad range of applications in design and analyses of experiments also in the wider biological community.

Keywords: experimental design, plant breeding, agriculture, LSD, HSD, Waller, Duncan, SNK, REGW, Scheffe, Durbin, Kruskal-Wallis, Friedman, Waerden, AMMI, diffograph, lattice, alpha design, PBIB, BIB, Index Smith, AUDPC, AUDPS, biodiversity, non-parametric analysis

INTRODUCTION

Computational protocols to analyze field experiments are an important part in breeding and agronomic experiments. The equal access to this tools enables communities of practices beyond institutional and country boundaries; it is therefore important in trait observation networks and decentralized breeding programs as is the case at the International Potato Center (CIP). In the early 2000s partners asked to use certain statistical analysis protocols but would have had to buy the basic commercial software. On the other hand, at this time the concept of open source and free software for statistics had become more visible and so we decided to use the R software (R-Core-Team, 2015) as a platform to disseminate in-house tools and protocols to the wider community in the spirit of free academic exchange and the production of global public goods. Initially, these were compiled as simple scripts. It turned out that at the time there was no R package available for design and analysis of agricultural and plant breeding experiments. Therefore we eventually decided to convert the scripts into a package. We named the package *agricolae* which is Latin for *(dedicated) to the farmer* - as per CIP's vision and mission (`http://cipotato.org/about-cip/vision-mission-values/`).

Over a course of a decade the package *agricolae* has been constantly revised and updated thanks to extensive feedback from users both in-house, nationally and around the world. Initial versions were presented at a R user conference (Mendiburu and Simon, 2007) and the International Society for Tropical Root Crops (ISTRC) conference at CIP headquarters in Lima (Mendiburu and Simon, 2009). In addition, this package was presented as a thesis subject for a Masters degree of the first author (`http://tarwi.lamolina.edu.pe/~fmendiburu/`. We wrote this article as more and more original research is being included in the package and the usage broadens. Here we present an overview of the current status, a brief review of reception and provide an outlook into the next versions.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1404v1 | CC-BY 4.0 Open Access | rec: 29 Sep 2015, publ: 29 Sep 2015

1 MATERIALS AND METHODS

1.1 Package development

The functions in the package are mainly based on previously published work. According to the best practices enforced by R (R-Core-Team, 2015) package development guidelines all functions are documented with the corresponding citations and examples of usage. Where ever possible, methods were tested against reference data sets and results compared against results reported in textbooks and from alternative software - mainly using SAS software (SAS-Institute, 2015). Many functions are documented with the textbook data sets or slightly modified real word data sets. Some functions were recently added reflecting the first authors own statistical research. These will be reported in more detail separately but briefly introduced here. News on latest additions are maintained on the first author's (FM) web-site (`http://tarwi.lamolina.edu.pe/~fmendiburu/`).

1.2 Usage analysis

We combine several means to assess uptake of the package in the community. Unfortunately, the R web-site itself does not provide any tracking mechanism for usage of packages. Therefore, we present here a few alternative approaches to quickly assess the utility of the package. A qualitative survey on user satisfaction was conducted as part of (Mendiburu, 2009). It is based on answers from 35 Peruvian and 13 international respondents. We also searched Google Scholar (`http://scholar.google.com` (accessed on Sep., 20th, 2015) for references to *agricolae* using the search term **agricolae Mendiburu -[citation]** - the latter exclusion word was necessary to remove references to R community servers. We filtered the search by year and reviewed the list manually to exclude any obvious non-relevant references. Lastly, some collaborators communicated their usage of *agricolae* as part of bigger projects personally (to RS).

2 RESULTS

2.1 Package development

Development started out in 2004 (Mendiburu and Simon, 2009), about four years after the first release of R in February, 2000. The first version of *agricolae* was released on CRAN in 2006 (see: `https://cran.r-project.org/src/contrib/Archive/agricolae/`). From 2006 till now (2015) 21 versions were released. As of September, 2015, the latest published version of *agricolae* is 1.2-2 `https://cran.r-project.org/web/packages/agricolae/index.html`. Here we summarize highlights by group and add usage examples; more details on usage can be obtained from the package documentation and the included tutorial. The package contains 35 reference datasets and 86 functions as per the up-coming package version 1.2-3 (scheduled for October, 2015). We present here already main new additions and improvements.

2.2 Descriptive statistics

This is a group of functions complementary to standard R functions and provides some convenience to produce graphs typically used in reports in the agricultural and breeding communities. More specifically, it contains the function *graph.freq()* to produce custom histograms. Other custom functions produce group wise statistics including histograms.

This group also includes other helper and convenience functions such as *montecarlo(), correlation(), tapply.stat(), skewness(), kurtosis(), and waller()*. The availability of function *skewness()* is the reason why *agricolae* is listed in the R taskview on *Distributions.*

Below we include the code to produce a histogram on grouped data and the resulting Figure 1.

```
 # Know your parameters
 str(design.rcbd)

## function (trt, r, serie = 2, seed = 0, kinds = "Super-Duper",
## first = TRUE,
## continue = FALSE)

 trt <- c("A", "B", "C","D","E")
 repeticion <- 4
 outdesign <- design.rcbd(trt,r=repeticion,
                          seed=-513,
                          serie=2 # a numbering scheme for labels
                          )
 book2<- zigzag(outdesign) # zigzag numeration
 print(t(matrix(book2$plots,5,4)))
```

```
##      [,1] [,2] [,3] [,4] [,5]
##      [1,]  101  102  103  104  105
##      [2,]  205  204  203  202  201
##      [3,]  301  302  303  304  305
##      [4,]  405  404  403  402  401
```

Another code fragment and graph shows the utility of a small function for constructing a modified histogram with an overlaid polygon as in Figure 2.

```
library(agricolae)
x<-seq(10,40,5)
y<-c(2,6,8,7,3,4)
# Poligon and frecuency
h<-graph.freq(x,counts=y,col=colors()[86],xlab=" ", ylab="Frequency",
axes=FALSE)
axis(1,x,las=2)
axis(2,0:10)
polygon.freq(h,col="red",xlab="  ", ylab="")
title( main="Histogram and polygon", xlab="Variable X")
```

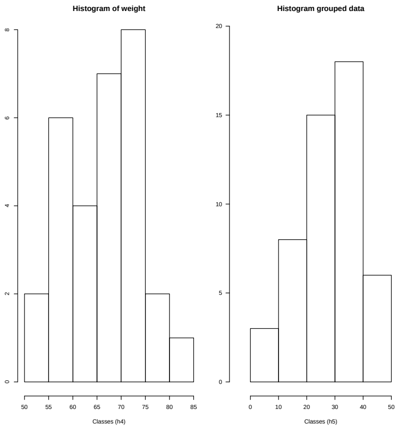

Figure 1. Histogram of grouped data produced by *agricoale*.

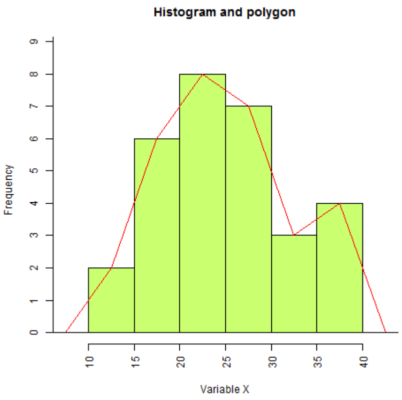

Figure 2. Histogram with an overlaid polygon produced by *agricoale*.

2.3 Experimental design

This group of 13 functions is at the core of this package. Their implementation is guided by (Cochran and Cox, 1992), Kuehl (2000), (Le Clerg, 1992), and (Montgomery, 2002). For individual references for each design function please refer to documentation in the package itself. These (and corresponding analysis functions as described in the next section) are the main features also described in the R taskview on *Experimental design* (`https://cran.r-project.org/web/views/ExperimentalDesign.html` where it also qualifies as a *core* package.

It contains most of the designs typically used in agricultural research like randomized complete block design, strip-plot design, split-plot design and factorial designs. Early on, modern designs like cyclic designs, balanced incomplete designs, alpha designs and augmented block designs were included. Other designs include the latin square design, the incomplete latin square design (the Youden design, a recent addition), the graeco-latin design, the completely randomized design, and the lattice design. The parameters of most functions have been standardized in naming and sequence, as well as their return value. The parameters include the specification of the randomization algorithm and the randomization seed thereby facilitating reproducibility. The return value is currently a list containing information about the parameters used, the resulting field book, summary statistics on the design like efficiency index, and a sketch showing the distribution of plots in the field. The companion function *zigzag()* allows to adjust the numbering of plots to the way a data collector would walk in the field. Two main improvements in the recent versions were: a) all design.* functions have gained a new parameter *randomization* with value TRUE or FALSE allowing the user now to completely turn off any randomization. b) the balanced incomplete block design has been fully optimized creating a design with a minimal number of blocks; it also returns a more friendly feedback if parameter combinations are not valid.

In the following example we show a simple case of a randomized complete block design.

```
library(agricolae)
# 4 treatments and 5 blocks
trt<-c("A","B","C","D")
outdesign <-design.rcbd(trt,5,serie=2,45,"Super-Duper") # seed = 45
rcbd <- outdesign$book
head(rcbd) # field book

##   plots block trt
## 1   101     1   C
## 2   102     1   B
## 3   103     1   D
## 4   104     1   A
## 5   201     2   B
## 6   202     2   D
```

Several design options are accompanied by complementary analysis functions; see the next section for a current list.

Here is an example of how to create an alpha design.

```
 Genotype<-paste("geno",1:30,sep="")
 r<-2
 k<-3
 plan_alpha <- design.alpha(Genotype,k,r,seed=5)

##
## alpha design (0,1) - Serie  I
##
## Parameters Alpha design
## =======================
## treatmeans : 30
## Block size : 3
## Blocks     : 10
## Replication: 2
##
## Efficiency factor
## (E ) 0.6170213
##
## <<< Book >>>

 2

## [1] 2
```

2.4 Multiple parametric comparisons

This group of functions include the Least Significant Difference test (*LSD.test()*), the Honestly Significant Difference test (*HSD.test()*), the Duncan test (*duncan.test()*, the Scheffe test (*scheffe.test()*, the Waller test (*waller.test()*, and the Student-Newman-Keuls test (*SNK.test()*). Their respective reference is: (Steel and Dickey, 1997).

A new test added in the up-coming version 1.2-3 is the Ryan, Einot and Gabriel and Welsch (REGW) test *REGW.test()* (Hsu, 1996). This test is preferable over other tests like Bonferroni or Duncan since it is more stringent. Another new feature is the calculation of p-values and confidence intervals for the following tests: *duncan.test(), SNK.test(), and REGW.test()*.

The list of six functions specifically designed for specialized ANOVA of experimental designs consists of: *BIB.test()* for analyzing the Balanced Incomplete Design, *DAU.test()* for Augmented Block Design analysis, *PBIB.test()* for Partially Balanced Incomplete Block Design, *sp.plot()* for split-plot analysis, *ssp.plot()* for split-split-plot analysis, and *strip.plot()* for strip-plot analysis.

As an illustration we show here the application of *PBIB.test()* to the previous example data of an alpha design.

```
 yield <-c(5,2,7,6,4,9,7,6,7,9,6,2,1,1,3,
           2,4,6,7,9,8,7,6,4,3,2,2,1,1,2,
           1,1,2,4,5,6,7,8,6,5,4,3,1,1,2,
           5,4,2,7,6,6,5,6,4,5,7,6,5,5,4)

 data<-data.frame(plan_alpha$book,yield)
 rm(yield,Genotype)

  # The analysis:
 attach(data)
 modelPBIB <- PBIB.test(block, Genotype, replication,
                        yield, k=3, group=TRUE,
 console=TRUE)

##
## ANALYSIS PBIB:  yield
##
## Class level information
## block : 20
## Genotype : 30
##
## Number of observations:  60
##
## Estimation Method:  Residual (restricted) maximum likelihood
##
## Parameter Estimates
##                       Variance
## block:replication 2.834033e+00
## replication       8.045902e-09
## Residual          2.003098e+00
##
##                       Fit Statistics
## AIC                        213.65937
## BIC                        259.89888
## -2 Res Log Likelihood      -73.82968
##
## Analysis of Variance Table
##
## Response: yield
##            Df Sum Sq Mean Sq F value Pr(>F)
## Genotype  29 72.006  2.4830  1.2396 0.3668
## Residuals 11 22.034  2.0031
##
## coefficient of variation: 31.2 %
## yield Means: 4.533333
##
## Parameters PBIB
##                       .
## Genotype             30
## block size            3
## block/replication 10
## replication           2
##
## Efficiency factor 0.6170213
##
## Comparison test lsd
##
## <<< to see the objects: means, comparison and groups. >>>

 detach(data)
```

2.5 Multiple non-parametric comparisons

The available tests for multiple non-parametric comparisons (Montgomery, 2002) are: *kruskal(), waerden.test(), friedman() and durbin.test()*. An **important note**to avoid possible confusion: R itself provides the functions *kruskal.test() and friedman.test()* - these report only a single value; whereas the functions provided by *agricolae* perform multiple comparisons.

Specific popular applications are the *friedman.test()* in participatory variety selection, for example in organo-leptic trials in order to judge quality of marketable plant parts. Below is an example of the latter.

```
library(agricolae)
data(grass)
attach(grass)
out<-friedman(judge,trt, evaluation,alpha=0.05, group=TRUE,console=TRUE,
main="Data of the book of Conover")

##
## Study: Data of the book of Conover
##
## trt,  Sum of the ranks
##
##    evaluation  r
## t1       38.0 12
## t2       23.5 12
## t3       24.5 12
## t4       34.0 12
##
## Friedman's Test
## ===============
## Adjusted for ties
## Value: 8.097345
## Pvalue chisq : 0.04404214
## F value : 3.192198
## Pvalue F: 0.03621547
##
## Alpha     : 0.05
## t-Student : 2.034515
## LSD       : 11.48168
##
## Means with the same letter are not significantly different.
## GroupTreatment and Sum of the ranks
## a    t1    38
## ab    t4    34
## b    t3    24.5
## b    t2    23.5

detach(grass)
```

2.6 Graphs for multiple comparisons

The two principal functions here are: *bar.group() and bar.err()*. We show a short code example here and the corresponding graphs in Figure 3 and Figure 4.

An up-coming addition is the diffograph (Figure 5 following (Hsu, 1996)). It provides an alternative way of showing differences between genotypes. Lines crossing the diagonal in this chart are equivalent to non-significant differences between genotypes.

```
library(agricolae)
data(sweetpotato)
model<-aov(yield~virus,data=sweetpotato)
out <- waller.test(model,"virus", console=TRUE,
main="Yield of sweetpotato\ndealt with different virus")

bar.err(out$means,variation="SE",horiz=TRUE,xlim=c(0,45),
        bar=FALSE,col=0)
```

```
bar.err(out$means,variation="range",ylim=c(0,45),bar=FALSE,col="grey",
         main="range")
```

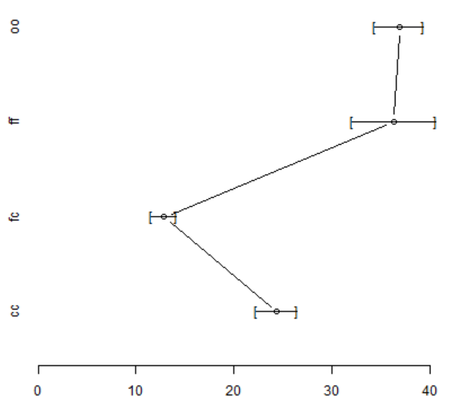

Figure 3. Error line chart.

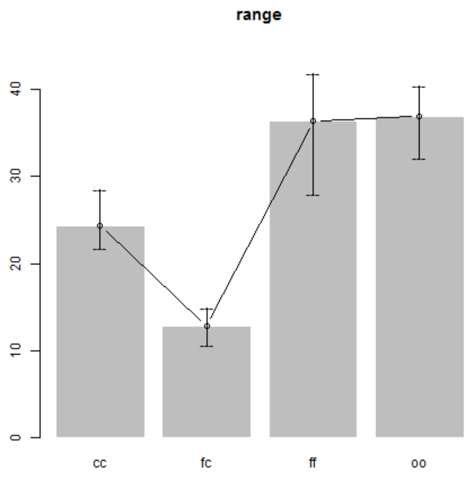

Figure 4. Error bar chart.

2.7 Stability analysis

Stability analysis in breeding refers to the simultaneous selection for yield and stability across different environments. Again, two functions are available for parametric and non-parametric cases. The parametric

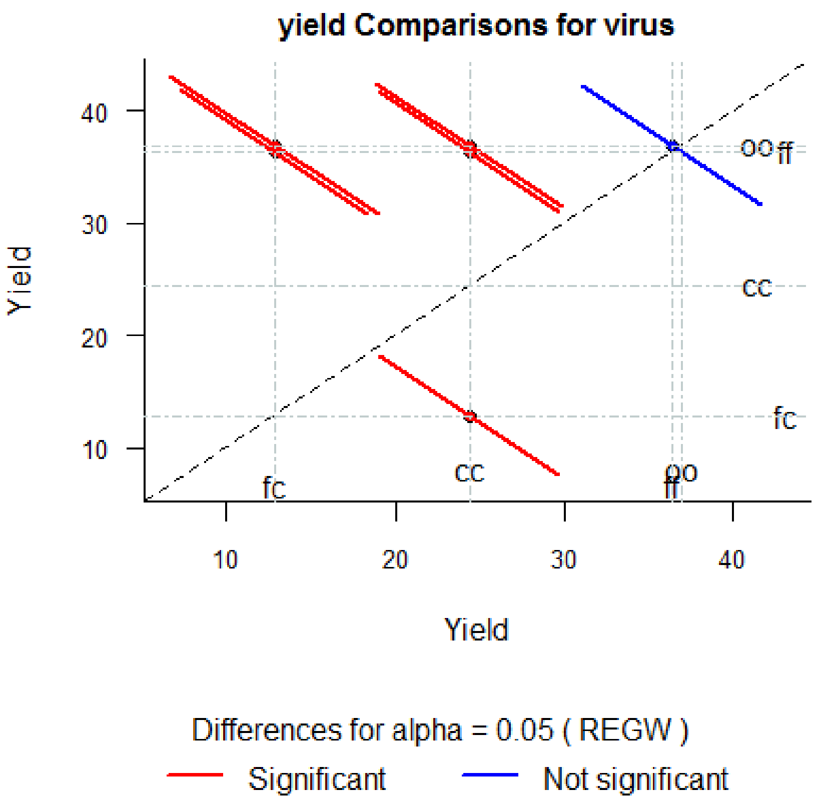

Figure 5. A diffograph for multiple comparisons following (Hsu, 1996). Lines are proportional to differences between genotypes. Lines crossing the diagonal correspond to non-significant differences.

stability function *stability.par()* implements Shukla's and Kang's stability variance (Kang, 1993). The non-parametric stability function *stability.nonpar()* implements Hayne's method (Haynes, 2009).

The stability analysis protocol should be further extended when there is significant interaction between genotype and environment (Crossa, 1990). Significant interaction exists when the first two component terms of a principal component analysis sum to more than 50%. In this case, the biplot visualization can be used; or also a triplot analysis for the first three components. We provide the function *plot.AMMI()* to visualize the interaction. The parameter *type* specifies if a biplot (Figure 6) or triplot (Figure 7) is created.

This same parameter has recently gained another addition to create a third type of graph - the *influence of genotype* graph. This is based on the realization that principal components represent an equi-distant representation and allow the application of spatial procedures. The end result (Figure 8 shows the relatedness of genotypes based on their multidimensional similarity. The details of this procedure will be published separately.

Here we elaborate an example of a parametric case, followed by an AMMI analysis.

```
##AMMI
options(digit=2)
v1 <- c(10.2,8.8,8.8,9.3,9.6,7.2,8.4,9.6,7.9,10,9.3,8.0,10.1,9.4,10.8,
6.3,7.4)
v2 <- c(7,7.8,7.0,6.9,7,8.3,7.4,6.5,6.8,7.9,7.3,6.8,8.1,7.1,7.1,6.4,
4.1)
v3 <- c(5.3,4.4,5.3,4.4,5.5,4.6,6.2,6.0,6.5,5.3,5.7,4.4,4.2,5.6,5.8,
3.9,3.8)
## 55

## [1] 55
```

```
v4 <- c(7.8,5.9,7.3,5.9,7.8,6.3,7.9,7.5,7.6,5.4,5.6,7.8,6.5,8.1,7.5,
5.0,5.4)
v5 <-c(9,9.2,8.8,10.6,8.3,9.3,9.6,8.8,7.9,9.1,7.7,9.5,9.4,9.4,10.3,
8.8,8.7)
study <- data.frame(v1, v2, v3, v4, v5)
rownames(study) <- LETTERS[1:17]
output <- stability.par(study, rep=4, MSerror=2)

altitude<-c(1200, 1300, 800, 1600, 2400)
stability <- stability.par(study,rep=4,MSerror=2, cova=TRUE,
name.cov= "altitude",
file.cov=altitude)

rdto <- c(study[,1], study[,2], study[,3], study[,4], study[,5])
environment <- gl(5,17)
genotype <- rep(rownames(study),5)
model<-AMMI(ENV=environment, GEN=genotype, REP=4, Y=rdto, MSE=2,
console=TRUE)

##
## ANALYSIS AMMI:  rdto
## Class level information
##
## ENV:  1 2 3 4 5
## GEN:  A B C D E F G H I J K L M N O P Q
## REP:  4
##
## Number of means:  85
##
## Dependent Variable: rdto
##
## Analysis of variance
##            Df    Sum Sq    Mean Sq F value        Pr(>F)
## ENV         4 734.2475 183.561882
## REP(ENV)   15
## GEN        16 120.0875   7.505471 3.752735 3.406054e-06
## ENV:GEN    64 181.2725   2.832382 1.416191 3.279630e-02
## Residuals 240 480.0000   2.000000
##
## Coeff var  Mean rdto
## 19.16584   7.378824
##
## Analysis
##     percent  acum Df    Sum.Sq  Mean.Sq F.value    Pr.F
## PC1    38.0  38.0 19 68.96258 3.629609    1.81 0.0225
## PC2    29.8  67.8 17 54.02864 3.178155    1.59 0.0675
## PC3    22.5  90.4 15 40.84756 2.723170    1.36 0.1680
## PC4     9.6 100.0 13 17.43370 1.341054    0.67 0.7915

pc <- model$analysis[, 1]
pc12<-sum(pc[1:2])
pc123<-sum(pc[1:3])
rm(rdto,environment,genotype)

plot(model,type=1,las=1)
plot(model,type=2,las=1)
```

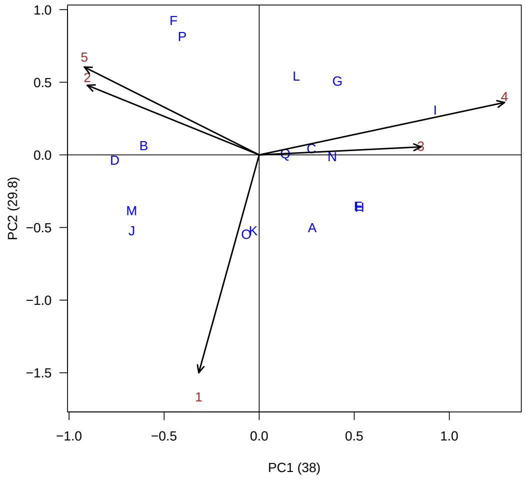

Figure 6. An illustrative biplot graph produced by *agricolae*.

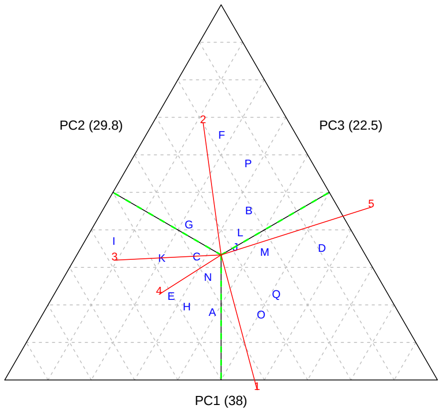

Figure 7. An illustrative triplot graph produced by *agricolae*.

Another related visualization is the addition of a contour line which adds a contour line proportional to the longest distance of a genotype and has a values between 0 and 1 (Figure 9).

```
library(agricolae)
# see AMMI.
data(sinRepAmmi)
Environment <- sinRepAmmi$ENV
Genotype <- sinRepAmmi$GEN
Yield <- sinRepAmmi$YLD
REP <- 3
MSerror <- 93.24224
model<-AMMI(Environment, Genotype, REP, Yield, MSerror)

# First call the plot function, then AMMI.contour()
plot(model)
AMMI.contour(model,distance=0.7,shape=8,col="red",lwd=2,lty=5)
```

Based on the AMMI analysis we can calculate the AMMI stability value (ASV) and the Yield stability value (YSV) using the function *index.AMMI()* following (Purchase, 1997) and (Sabaghnia and Dehghani, 2008).

```
index<-index.AMMI(model)
# Crops with improved stability according AMMI.
print(index[order(index[,3]),])

##           ASV  YSI  rASV  rYSI  means
## Q 0.1559316   18     1    17   5.88
## C 0.3138851   11     2     9   7.44
## N 0.4347076    5     3     2   7.92
## K 0.5205310   19     4    15   7.12
## O 0.5494696    6     5     1   8.30
## (shortened)
```

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1404v1 | CC-BY 4.0 Open Access | rec: 29 Sep 2015, publ: 29 Sep 2015

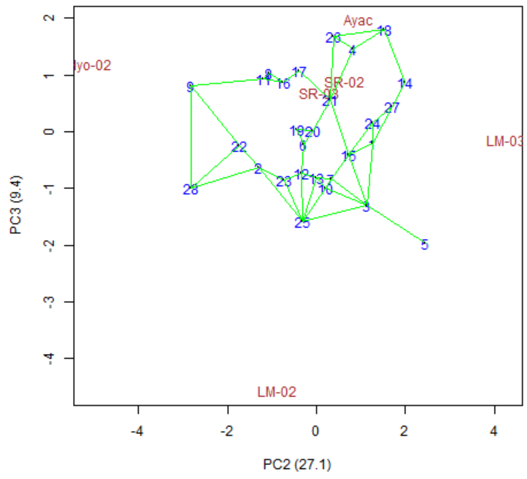

Figure 8. An illustrative 'influence of genotype' graph

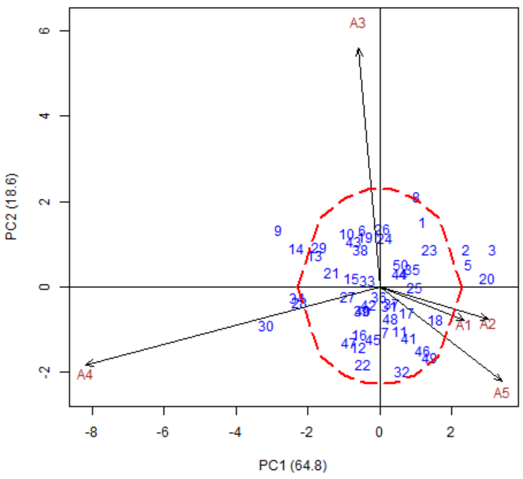

Figure 9. An illustrative AMMI contour graph

```
# Crops with better response and improved stability according AMMI.
print(index[order(index[,4]),])

##           ASV YSI rASV rYSI means
## O 0.5494696   6    5    1  8.30
## N 0.4347076   5    3    2  7.92
## G 0.6905095  13   10    3  7.90
## A 0.5909737  11    7    4  7.86
## H 0.6959572  16   11    5  7.68
## (shortened)
```

2.8 Quality of field experiments

An important assumption in field experiments is the homogeneity of soil across plots. To test this hypothesis the Smith index (Gomez and Gomez, 1976) has been developed and has been implemented in the function *index.smith()* (Figure 10).

An indicator of field experiment quality is the coefficient of variation. It is easily calculated using the function *cv.model()*.

If there are doubts about the adequateness of a simple model this can be tested for non-additive effects. Based on Tukey's test, the function *nonadditivity()* can here be of use.

When there is an indication that errors have a non-normal distribution in a variance analysis, the probability and significance of different sources can be calculated using *resampling.model()*.

The function *simulation.model()* helps to assess the proportion of valid results of an ANOVA by generating pseudo-experimental errors under the assumption of normality.

2.9 Genetic designs and their analysis

Genetic designs here are mating designs used in breeding. These designs originated at the North Carolina State University (NCSU) and comprise three designs. All of them can be created via the single function *carolina()*. The function *lineXtester()* can be applied to analyze the data obtained for these designs. The main reference is (Singh, 1979).

An example is listed here:

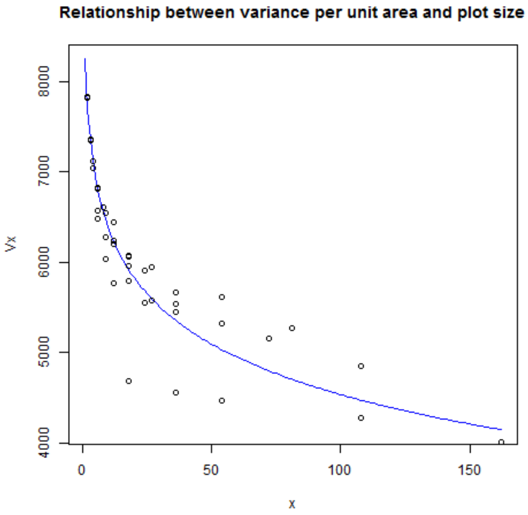

Figure 10. An illustrative figure based on *index.smith()*.

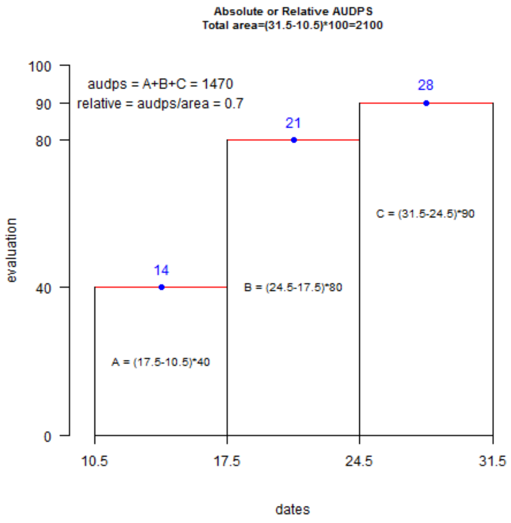

Figure 11. An illustrative figure based on *audps()*.

```
library(agricolae)
# example 1
data(heterosis)
names(heterosis)[2:3] = c("Rep", "Treat")
head(heterosis)

##   Place Rep Treat   Factor Female   Male    v1   v2    v3    v4     v5
## 1     1   1     1 progenie   LT-8  TS-15 0.948 1.65 17.22  9.93 102.58
## 2     1   1     2 progenie   LT-8 TPS-13 1.052 2.20 17.84 12.45 107.37
## 3     1   1     3 progenie   LT-8 TPS-67 1.050 1.88 15.61  9.30 120.49
## 4     1   1     4 progenie  TPS-2  TS-15 1.058 2.00 16.04 12.77  83.78
## 5     1   1     5 progenie  TPS-2 TPS-13 1.123 2.45 16.48 14.13  90.40
## 6     1   1     6 progenie  TPS-2 TPS-67 1.115 2.63 18.72 14.60  81.79

site1<-subset(heterosis,heterosis[,1]==1)
attach(site1)
output1<-lineXtester(Rep, Female, Male, v2)

##
## ANALYSIS LINE x TESTER:  v2
##
## ANOVA with parents and crosses
## ==============================
##                          Df       Sum Sq     Mean Sq F value Pr(>F)
## Replications              2  0.002674074 0.001337037   0.088 0.9159
## Treatments               35 29.135740741 0.832449735  54.763 0.0000
## Parents                  11 21.221688889 1.929244444 126.917 0.0000
## Parents vs. Crosses       1  1.692474074 1.692474074 111.341 0.0000
## Crosses                  23  6.221577778 0.270503382  17.795 0.0000
## Error                    70  1.064059259 0.015200847
## Total                   107 30.202474074
##
## ANOVA for line X tester analysis
## ================================
##                 Df    Sum Sq    Mean Sq F value Pr(>F)
## Lines            7 4.3497111 0.62138730   9.078 0.0003
## Testers          2 0.9135444 0.45677222   6.673 0.0092
## Lines X Testers 14 0.9583222 0.06845159   4.503 0.0000
## Error           70 1.0640593 0.01520085
##
## ANOVA for line X tester analysis including parents
## ==================================================
##                       Df       Sum Sq     Mean Sq F value Pr(>F)
## Replications          2  0.002674074 0.001337037   0.088 0.9159
## Treatments           35 29.135740741 0.832449735  54.763 0.0000
## Parents              11 21.221688889 1.929244444 126.917 0.0000
## Parents vs. Crosses   1  1.692474074 1.692474074 111.341 0.0000
## Crosses              23  6.221577778 0.270503382  17.795 0.0000
## Lines                 7  4.349711111 0.621387302   9.078 0.0003
## Testers               2  0.913544444 0.456772222   6.673 0.0092
## Lines X Testers      14  0.958322222 0.068451587   4.503 0.0000
## Error                70  1.064059259 0.015200847
## Total               107 30.202474074
##
## GCA Effects:
## ===========
## Lines Effects:
## Achirana     LT-8     MF-I    MF-II  Serrana    TPS-2   TPS-25    TPS-7
##    0.143   -0.346   -0.134   -0.169    0.394   -0.016    0.321   -0.194
##
## Testers Effects:
## TPS-13 TPS-67  TS-15
##  0.104  0.052 -0.156
##
## SCA Effects:
## ===========
##           Testers
## Lines      TPS-13 TPS-67  TS-15
##   Achirana -0.010  0.002  0.008
##   LT-8      0.186 -0.042 -0.144
##   MF-I     -0.059  0.070 -0.011
##   MF-II    -0.021  0.055 -0.034
##   Serrana  -0.241  0.008  0.233
##   TPS-2    -0.044  0.101 -0.057
##   TPS-25    0.136 -0.259  0.123
##   TPS-7     0.054  0.064 -0.118
##
## Standard Errors for Combining Ability Effects:
## =============================================
## S.E. (gca for line)   : 0.04109724
## S.E. (gca for tester) : 0.02516682
## S.E. (sca effect)     : 0.0711825
## S.E. (gi - gj)line    : 0.05812027
## S.E. (gi - gj)tester  : 0.03559125
## S.E. (sij - skl)tester: 0.1006673
##
## Genetic Components:
## ==================
## Cov H.S. (line)   : 0.0614373
```

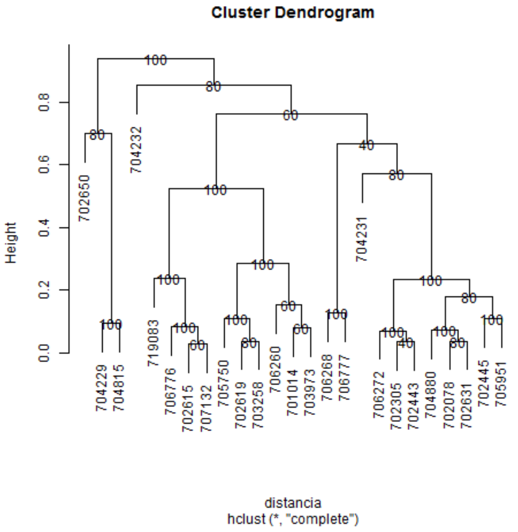

Figure 12. An illustrative figure based on *consensus()*.

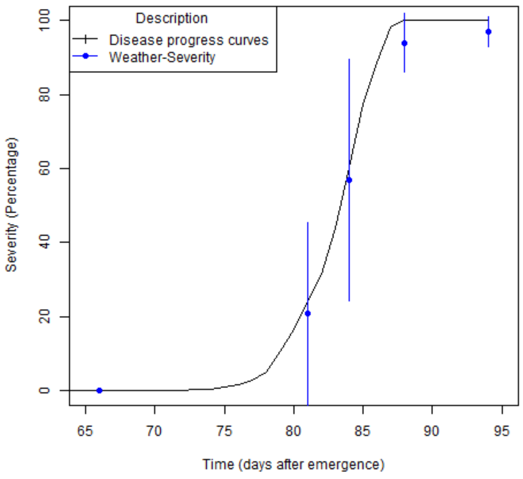

Figure 13. An illustrative figure based on *lateblight()*.

```
## Cov H.S. (tester) : 0.01618003
## Cov H.S. (average): 0.004651843
## Cov F.S. (average): 0.1223343
## F = 0, Adittive genetic variance: 0.01860737
## F = 1, Adittive genetic variance: 0.009303686
## F = 0, Variance due to Dominance: 0.03550049
## F = 1, Variance due to Dominance: 0.01775025
##
## Proportional contribution of lines, testers
##  and their interactions to total variance
## ===========================================
## Contributions of lines  : 69.91331
## Contributions of testers: 14.68349
## Contributions of lxt    : 15.4032

detach(site1)
```

2.10 Other biological analyses

Here we just briefly mention tools useful in biodiversity analysis and genetic dendrogram verification. They include: *index.bio()* and *consensus()*. The function *index.bio()* calculates several indices (Magurran, 1988) and also their confidence intervals using bootstrap (Tibshirani, 1993). The function *consensus()* uses bootstrap to calculate relative frequencies (Figure 12). As a side effect, the function also filters out exact duplicates. It provides access to a variety of distances and methods; please refer to the package documentation for more details.

Similarly, the function *path.analysis()* is included for convenience and can help to elucidate complex relationships between biological co-variates and response variables.

A couple of functions (*audpc() and audps()* is useful for analyzing disease progress. A recent addition is *audps()* (Simko and Piepho, 2011). It improves upon the AUDPC approach by giving a weight closer to optimal to the first and last observations; see Figure 11.

A last very specific function models the disease curve of late blight *Phytophthora infestans*. *lateblight()* simulates the effect of weather, host growth and resistance, and fungicide use on asexual development and growth of this oomycete on potato foliage (Figure 13). For detailed references see the package documentation.

2.11 Usage and reception

A first result of being part of the R package developer's community was it's external feedback. Early on, *agricolae* was prominently mentioned on the R task view section on 'Experimental Design' `https://cran.r-project.org/web/views/ExperimentalDesign.html` and listed there as a 'core' package since about 2008.

The main result from the user survey was that users were satisfied as indicated by an Ikert index of 0.81.

The Google Scholar data indicate that the package is increasingly accepted and cited as indicated by the exponential growth curve in Fig. 14. Currently, a total of 179 citations are reported by Google Scholar. Examples of usage from an incomplete list include in horticulture (Merk et al., 2012), (George et al., 2014); in biological conservation (Giroldo and Scariot, 2015); in cancer research: (Martınez et al., 2013); in aquaculture: (Bergström et al., 2015); in ecosystems and environment (Schwab et al., 2015).

As can be seen in Fig. 15 the package has also had an increasing trend of installations over the past three years. The absolute numbers are only indicative since the counts are only registered via one download server. Also, the records for 2012 reflect only the last quarter of that year, whereas for 2015 the downloads for the last quarter are still missing as per time of writing (September, 2015). The total reported here are 77447 downloads.

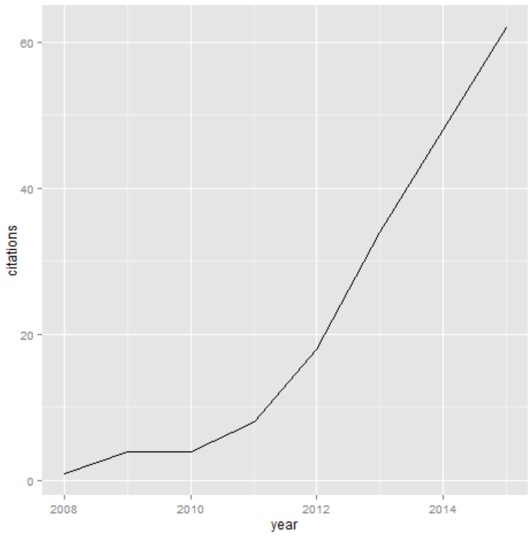

Figure 14. Agricolae citations as registered through Google Scholar.

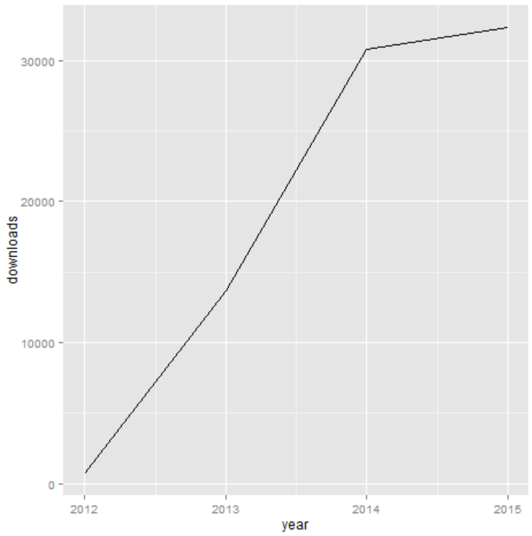

Figure 15. Agricolae downloads as registered through RStudio.

Lastly, an increasing list of developers working on breeding databases include *agricolae* in the backend - often as part of a toolchain to create fieldbooks. This list of breeding database includes to the best of our knowledge the: Integrated Breeding Platform (IBP, `https://www.integratedbreeding.net/`), the Cassavabase (`http://www.cassavabase.org/`, and the KDdart (`http://www.diversityarrays.com/kddart`).

3 DISCUSSION

As anticipated, the development of *agricolae* has filled a niche and has become a useful tool to encapsulate personal knowledge in an explicit and transparent way. It serves to share best practices in it's application domain with national collaborators and beyond. It is increasingly used beyond breeding or agricultural use cases. It seems the R platform (an example of an open source model) has facilitated knowledge transfer and sharing. The constant external review of packages by the central R maintainers as well as review by use also stimulates the regular updates including improvements to the tool and improvements of personal professional skills as statistician and programmer. Lastly, the package is increasingly used in the wider biological community testifying to it's general utility. In summary, the development of *agricolae* has been rewarding at a professional, institutional and community level.

agricolae along with data sets is available from the R community repository `https://cran.r-project.org/web/packages/agricolae/index.html`.

ACKNOWLEDGMENTS

We acknowledge the *agricolae* users for their feedback and suggestions. We particularly would like to thank colleagues from CIP for providing data and feedback.

REFERENCES

Bergström, P., Carlsson, M. S., Lindegarth, M., Petersen, J. K., Lindegarth, S., and Holmer, M. (2015). Testing the potential for improving quality of sediments impacted by mussel farms using bioturbating polychaete worms. *Aquaculture Research*.

Cochran and Cox (1992). *Experimental Design*. John Wiley and Sons, Inc, New York. Wiley Classics Library Edition.

Crossa, J. (1990). Statistical analysis of multilocation trials. *Advances in Agronomy*, 44:55–85.

George, N. A., Shankle, M., Main, J., Pecota, K. V., Arellano, C., and Yencho, G. C. (2014). Sweetpotato grown from root pieces displays a significant genotype× environment interaction and yield instability. *HortScience*, 49(8):984–990.

Giroldo, A. and Scariot, A. (2015). Land use and management affects the demography and conservation of an intensively harvested cerrado fruit tree species. *Biological Conservation*, 191:150–158.

Gomez, K. and Gomez, A. (1976). *Statistical Procedures for Agriculture Research. Second Edition*. John Wiley and Sons, NY, USA.

Haynes, e. a. (2009). Phenotypic stability of resistance to late blight in potato clones evaluated at eight sites in the united states of america. *Journal Potato Research*, 75:211–221.

Hsu, J. C. (1996). *Multiple comparisons theory and methods*. Chapman Hall/CRC. Departament of statistics the Ohio State University. USA.

Kang, M. S. (1993). Simultaneous selection for yield and stability: Consequences for growers. *Journal Potato Research*, 85:754–757.

Kuehl, R. (2000). *Design of Experiments*. Duxburry. 2nd edition.

Le Clerg, E. (1992). *Field plot technique*. Burgess Publishing Company.

Magurran, A. (1988). *Ecological diversity and its measurement*. Princeton University Press.

Martınez, M., Pan, J., Polya, D. A., and Giri, A. K. (2013). High arsenic in rice is associated with elevated genotoxic effects in humans. *Scientific reports*, 3.

Mendiburu and Simon (2007). Agricolae: A free statistical library for agricultural research. UseR conference.

Mendiburu and Simon (2009). Agricolae - a free statistical toolbox for agricultural experiments. ISTRC conference.

Mendiburu, F. (2009). Una herramienta de análisis estadístico para la investigación agrícola. tesis para optar el grado académico de maestro en ciencias con mención en ingeniería de sistemas.

Merk, H. L., Yarnes, S. C., Van Deynze, A., Tong, N., Menda, N., Mueller, L. A., Mutschler, M. A., Loewen, S. A., Myers, J. R., and Francis, D. M. (2012). Trait diversity and potential for selection indices based on variation among regionally adapted processing tomato germplasm. *Journal of the American Society for Horticultural Science*, 137(6):427–437.

Montgomery (2002). *Diseño y Análisis de Experimentos*. John Wiley and Sons, Inc, New York.

Purchase, J. (1997). Parametric analysis to describe genotype by environment interaction and yield stability in winter wheat.

R-Core-Team (2015). A language and environment for statistical computing.

Sabaghnia, S. and Dehghani (2008). The use of an ammi model and its parameters to analyse yield stability in multi-environment trials. *Journal of Agricultural Science*, 146:571–581.

SAS-Institute (2015). Sas software.

Schwab, N., Schickhoff, U., and Fischer, E. (2015). Transition to agroforestry significantly improves soil quality: A case study in the central mid-hills of nepal. *Agriculture, Ecosystems & Environment*, 205:57–69.

Simko, I. and Piepho, H.-P. (2011). The area under the disease progress stairs: Calculation, advantage, and application. *Analytical and Theoretical Plant Pathology*.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1404v1 | CC-BY 4.0 Open Access | rec: 29 Sep 2015, publ: 29 Sep 2015

Singh, C. (1979). *Biometrical Methods in Quantitative Genetic Analysis*. Kalyana, India.

Steel, T. and Dickey (1997). *Principles and Procedures of Statistic a Biometrical Approach.* Mc Graw Hill.

Tibshirani, R. (1993). *An Introduction to Bootstrap*. Chapman and Hall/CRC.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1404v1 | CC-BY 4.0 Open Access | rec: 29 Sep 2015, publ: 29 Sep 2015