Chapter 5
Advanced Multivariate and Computational Approaches in Agricultural Studies

Inayat Ur Rahman, Eduardo Soares Calixto, Aftab Afzal, Zafar Iqbal, Niaz Ali, Farhana Ijaz, Muzammil Shah, and Khalid Rehman Hakeem

Contents

5.1 Introduction 93
5.2 Methodology 94
5.3 Ordination Analyzes 95
5.4 Correlograms, Heatmaps, and Scatterplot Matrix 96
5.5 Violin and Box Plot 96
5.6 Chord Diagram and Bipartite Networks 99
5.7 Hierarchical Clustering 101
5.8 Final Remarks 101
References 102

5.1 Introduction

The statistical principles underlying design of experiments were pioneered by R. A. Fisher in the 1920s and 1930s at Rothamsted Experimental Station, an agricultural research station around 40 km north of London. Fisher had shown the way on how

I. U. Rahman (✉)
Department of Botany, Hazara University, Mansehra, Khyber Pakhtunkhwa, Pakistan

William L. Brown Center, Missouri Botanical Garden, St. Louis, MO, USA

E. S. Calixto
Department of Biology, University of São Paulo, São Paulo, Brazil

University of Missouri St. Louis (UMSL), Saint Louis, MO, USA

A. Afzal · Z. Iqbal · N. Ali · F. Ijaz
Department of Botany, Hazara University, Mansehra, Khyber Pakhtunkhwa, Pakistan

M. Shah
Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

K. R. Hakeem
Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

Princess Dr. Najla Bint Saud Al-Saud Center for Excellence Research in Biotechnology, King Abdulaziz University, Jeddah, Saudi Arabia

K. R. Hakeem et al. (eds.), *Essentials of Bioinformatics, Volume III*,
https://doi.org/10.1007/978-3-030-19318-8_5

to draw valid conclusions from field experiments where nuisance variables such as temperature, soil conditions, and rainfall are present. He had shown that the known nuisance variables usually cause systematic biases in results of experiments and the unknown nuisance variables usually cause random variability in the results and are called inherent variability or noise. He introduced the concept of analysis of variance (ANOVA) for partitioning the variation present in data due to (a) attributable factors and (b) chance factors. The methodologies he and his colleague Frank Yates developed are now widely used. No doubt, these methodologies have a profound impact on agricultural sciences research.

It may be emphasized in the beginning itself that experimental design is first about agriculture, animal science, biology, chemistry, industry, education, etc. and then about statistics and mathematics. In fact, experimental design forms the backbone of agricultural sciences; it is an integral component of every research endeavor in agricultural sciences. To design a good experiment, the researcher first needs to outline questions to be answered or needs one or more well-defined hypotheses.

Therefore, the application of advanced multivariate analyses and decision support tools or approaches are necessary and suggested so that researchers can better evaluate and present the results of different studies related to agricultural system. Thus the aim of this study is to provide a new support for analyses based on advanced multivariate approaches, concerning to the main statistical analyses used in agricultural system, as well as to provide better ways to show and manipulate the data graphically, enhancing the publication potential of the studies.

5.2 Methodology

Data used in this chapter were acquired from agricultural studies with the authors' authorization. For Sects. 5.4 and 5.5, we used data related to the chickpea (*Cicer arietinum* Medik) production. In this study, the authors analyzed the chickpea growth based on different soil treatments. The five treatments were (A) diammonium phosphate (DAP) half dose (12 g) + biofertilizer; (B) ammonium molybdate (0.236 g) + zinc sulphate (0.096 g) + biofertilizer; (C) ammonium molybdate (0.165 g) + zinc sulphate (0.144 g) + biofertilizer; (D) ammonium molybdate (0.236 g) + zinc sulphate (0.096 g); and (E) ammonium molybdate (0.165 g) + zinc sulphate (0.144 g). The parameters analyzed were plant height (PH) (cm), root length (RL) (cm), plant fresh weight (PFW) (g), plant dry weight (PDW) (g), root fresh weight (RFW) (g), number of flowers per plant (NoFPP), number of pods per plant (NoPPP), number of branches per plant (NoBPP), stem diameter (SD) (cm), and number of leaves per plant (NoLPP). For Sect. 5.6, we used data-related bands and genotypes.

All analyses and graphs elaborated were assembled in CANONO and RStudio 3.5.1 software at a level of 5% significance. The packages and main functions used in R are described in each topic addressed.

5.3 Ordination Analyzes

Overview Ordination techniques are used to describe relationships between dependent and independent variables. Principal Components Analysis (PCA) is one of the earliest ordination technique invented by Karl Pearson in 1901 (Dunn and Stearns 1987). Currently, it is mostly used as a tool in exploratory data analysis and for making predictive models. PCA uses a rigid rotation to derive orthogonal axes, which maximize the variance in the data set. Computationally, Principal components analysis is the basic eigen analysis technique. It maximizes the variance explained by each successive axis. The sum of the eigen values will equal the sum of the variance of all variables in the data set. PCA is relatively objective and provides a reasonable but crude indication of relationships i.e. in an indirect non-canonical way.

Example From PCA (Fig. 5.1), three groups, A–B, C–D, and E can be distinguished. These groups are separated by the analyzed characteristics of each treatment, where NoLPP (number of leaves per plant) and PH (plant height) seem to be the main factors that explain this grouping.

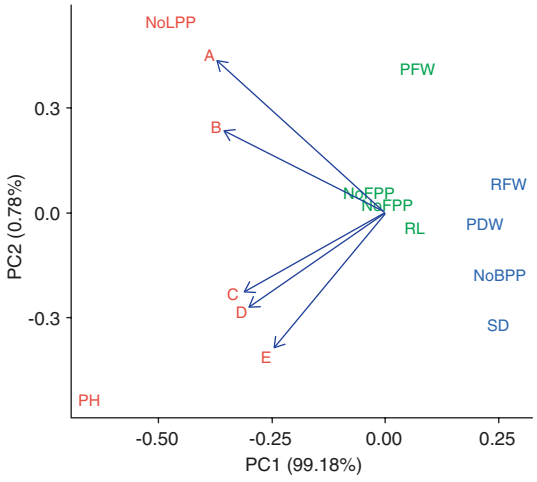

Fig. 5.1 PCA showing correlation of treatments in three groups (A–B, C–D, and E) based on the characteristics analyzed for each treatment, where NoLPP and PH seem to have a major influence in this grouping. *PH* plant height (cm), *RL* root length (cm), *PFW* plant fresh weight (g), *PDW* plant dry weight (g), *RFW* root fresh weight (g), *NoFPP* number of flowers per plant, *NoPPP* number of pods per plant, *NoBPP* number of branches per plant, *SD* stem diameter (cm), *NoLPP* number of leaves per plant. *Packages*: ggfortify (Tang et al. 2016), cluster (Maechler et al. 2018); *Functions*: autoplot, clara

5.4 Correlograms, Heatmaps, and Scatterplot Matrix

Overview Correlogram is a kind of correlation matrix, which shows the relationship between each pair of numerical variables analyzed based on the degree of association. Heatmap (Fig. 5.2) is a graphical tool based on a color-coding system to represent the relationship between pairs of variables and calculated by dissimilarity. A scatterplot matrix (Fig. 5.2) is a set of scatterplots organized in a matrix or grid and shows the relationship between pairs of variables. Scatterplot matrix is very useful for exploratory data analysis, especially for linear correlation between multiple variables.

Example From correlogram (Fig. 5.2, top left), it may be observed that in general there is a strong correlation among all parameters analyzed, except for SD. The lower part of the correlogram is the R^2 values, the diagonal is the parameter names, and the upper part is pie graph showing the same relation presented in the lower part. The color in both lower and upper parts represents the positive and negative relationship. In this example, we can see that blue color shows a positive relation, red color presents a negative relation, and white color represents no relation. The heatmap shows the relation among the parameters through distance of dissimilarity (in our case, Euclidian distance). Three different clusters may be seen in the heatmap of Fig. 5.2, top right. The first cluster, in green, is represented by treatments A and B (see Methodology for differences among treatment), the second is represented by treatments C and D, and the third is represented by treatment E. Thus, from the ten parameters analyzed, differences among treatments may be seen, and the chickpea growth has been strongly influenced by soil components (see Sect. 5.5). Lastly, similar to correlogram, the scatterplot (Fig. 5.2, bottom) shows the relation between two variables, where we can see the absolute value (correlation coefficient – "r") of the correlation between variables and the significance of the relationship discriminated by asterisks ($***p < 0.001$, $**p < 0.01$, $*p < 0.05$) on top, the bivariate scatterplots with fitted line on bottom, and histograms with kernel density estimation and rug plot on diagonal (Fig. 5.2, bottom). For instance, we can observe that SD presents a low correlation value with all other parameters.

5.5 Violin and Box Plot

Overview Box plot is a simple and standardized way to display data distribution. This plot is usually based on five elements: minimum, maximum, first and third quartile, and median. In addition, individual points can be plotted, especially if the data present outliers. This kind of graphic representation is very useful for analyzing

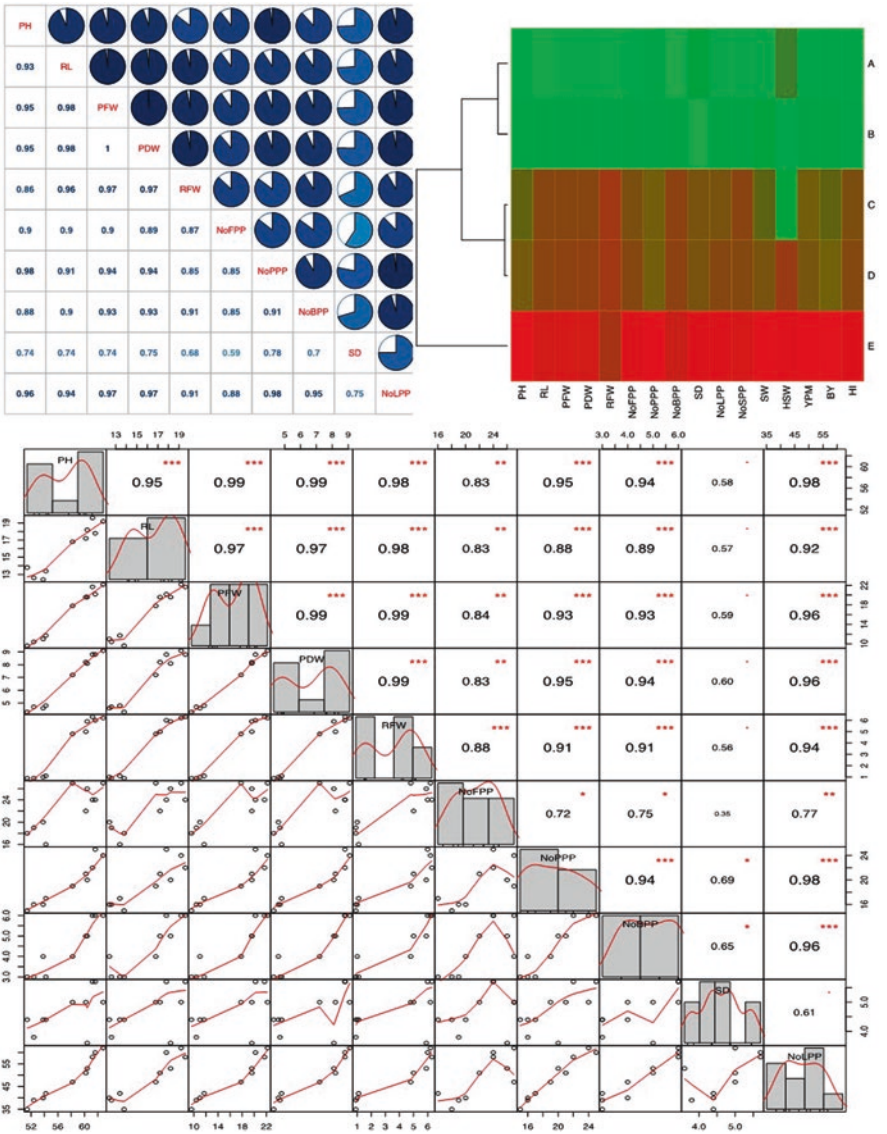

Fig. 5.2 Correlogram (top left), heatmap (top right), and scatterplot (bottom) of an agricultural data set based on chickpea (*Cicer arietinum*) growth in different soil moisture. *PH* plant height, *RL* root length, *PFW* plant fresh weight, *PDW* plant dry weight, *RFW* root fresh weight, *NoFPP* number of flowers per plant, *NoPPP* number of pods per plant, *NoBPP* number of branches per plant, *SD* stem diameter, *NoLPP* number of leaves per plant. For treatments A, B, C, D, and E, see Methodology. *Packages*: corrplot (Wei and Simko 2017), PerformanceAnalytics (Peterson and Carl 2018); *Functions*: coorplot, heatmap, chart.Correlation

variation in samples, for example, the degree of dispersion (spread) and skewness. Violin plot is a combination between box plot and density plot. In addition to all box plot features, violin plot shows the distribution shape of the data, that is, showing the kernel probability density of the data at different values, similar to histograms (see Rosenblatt 1956). Violin plot is specifically useful to see if the data is having multimodal (more than one peak) distribution or where the data have points that are more frequent.

Example The box plot (Fig. 5.3, left) shows variation in plant height among the treatments. In a general, we can observe that chickpea individuals that were seeded in treatments A and B were larger, with slightly higher values for treatment A. Similarly, plants with intermediate growth for treatments C and D and plants with low height for treatment E are indicated. The box plot with "notch" (Fig. 5.3, middle) is similar to box plot (Fig. 5.3, left), but it shows the confidence interval. Usually if we have overlap of notches between two groups, it means that there is no difference between them. Thus, comparing the treatments in Fig. 5.3 with notch, no differences between A and B or between C and D are observed. However, it may be assumed that there is a difference between A–B and C–D, A–B and E, and C–D and E. Furthermore, the violin plot (Fig. 5.3, right) shows the distribution of the data based on the kernel probability density. Analyzing the treatment B, we can see that there is no multimodal distribution of data, where most of the values are distributed around the median. On the other hand, treatment C, for instance, presents two peaks of distribution, showing a bimodal distribution.

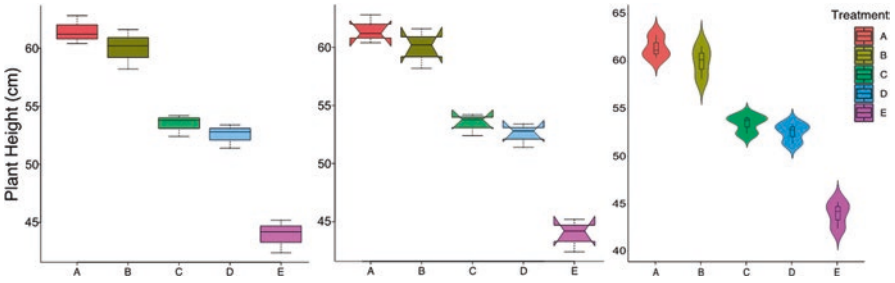

Fig. 5.3 Box plot (left), box plot with notch (middle), and violin plot (right) of an agricultural data set based on chickpea (*Cicer arietinum*) growth in different soil moisture. *PH* plant height, *RL* root length, *PFW* plant fresh weight, *PDW* plant dry weight, *RFW* root fresh weight, *NoFPP* number of flowers per plant, *NoPPP* number of pods per plant, *NoBPP* number of branches per plant, *SD* stem diameter, *NoLPP* number of leaves per plant. For treatments A, B, C, D, and E, see Methodology. *Package*: ggplot2 (Wickham 2016); *Function*: ggplot

5.6 Chord Diagram and Bipartite Networks

Overview Chord diagram is a graphical method to visualize the inter relationships between groups in a matrix, and it is very useful to see and compare connections, similarities, and differences among groups. Lines or arcs link one group to the other, and the width of the arc is proportional to the "importance" of the flow (if weighted). In a similar way, Bipartite Networks show connections (links) among nodes from two distinct sets and can be binary (presence/absence) or quantitative (weighted). In addition, different metrics (e.g., nestedness, connectance) can be used to evaluate the network. For this, one can use the *network-level* function in package "bipartite."

Example The chord diagram (Fig. 5.4, left) shows that the majority of bands present a high number of bindings with genotypes, except B1 and B2, which have links

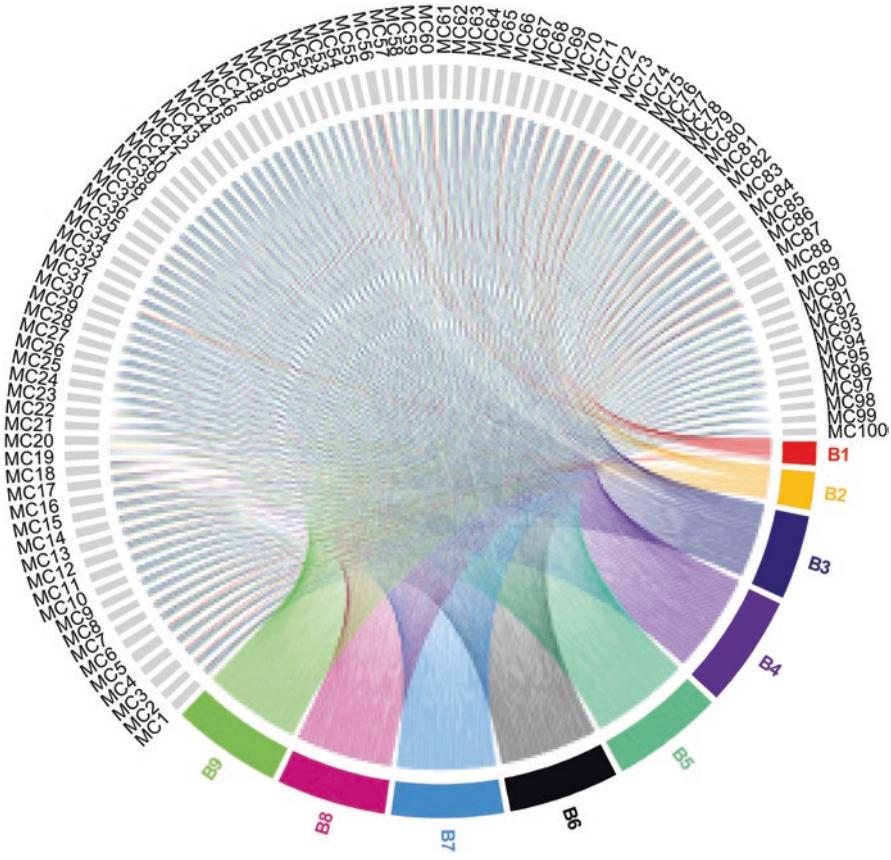

Fig. 5.4 Chord diagram showing the relationship between bands (B) and genotypes (MC)

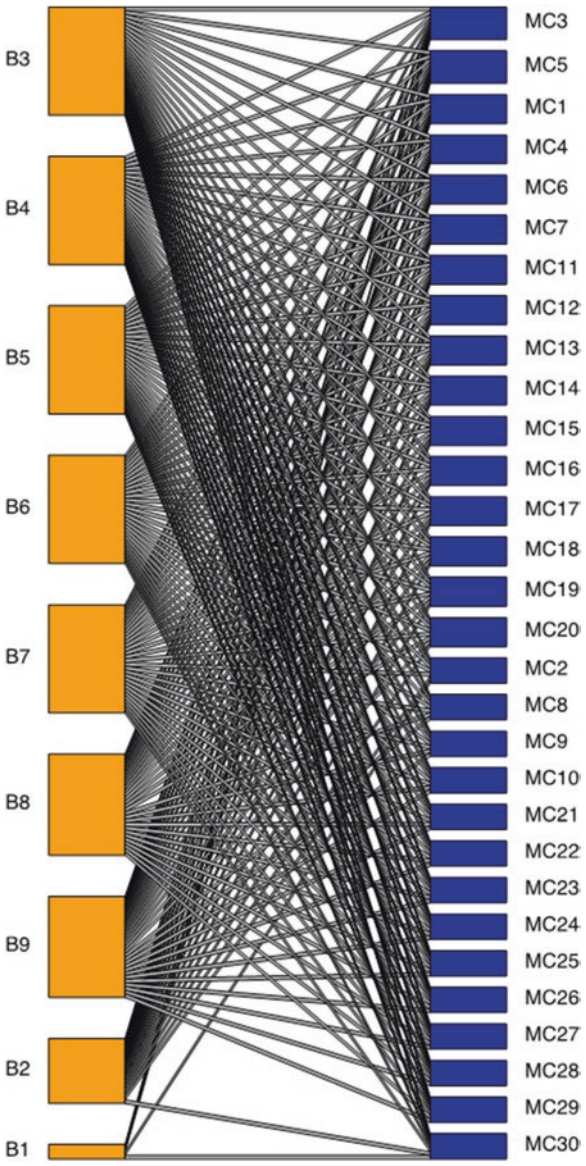

Fig. 5.5 Bipartite network showing the relationship between bands and genotypes in *plant species*. *B* bands (orange), *MC* genotype (blue). Observe the decreasing order of bands in relation to the number of connections with the genotypes. *Package*: circlize (Gu et al. 2014), bipartite (Dormann et al. 2008); *Functions*: chord diagram, plotweb

with only a few genotypes. Similarly, we can observe the same pattern of interactions in the bipartite network (Fig. 5.5). However, the network is more informative, since provides details of the decreasing order of interactions, showing the bands that have the greatest and the least number of interactions.

5.7 Hierarchical Clustering

Overview Hierarchical clustering or hierarchical cluster analysis is an algorithm that clusters similar objects into specific groups. These clusters have a predetermined ordering from top to bottom from distance measures, such as Euclidean distance.

Example The cluster dendrogram, based on Euclidean distance (Fig. 5.6), shows the grouping of treatments according to the characteristics of each group. In our case, three groups, A–B, C–D, and E are obsorved, which are in agreement with the graphics made in Sect. 5.5. Thus it may be suggested that there are differences between these three groups based on the Euclidean distance of the characteristics analyzed in the treatments.

5.8 Final Remarks

Many research studies in agricultural systems have been developed so far. However, the studies did not present adequate statistical approaches or graphical methods that explain and better represent the results obtained. This directly influences the visualization and dissemination of the works, and limits the scope and visibility of

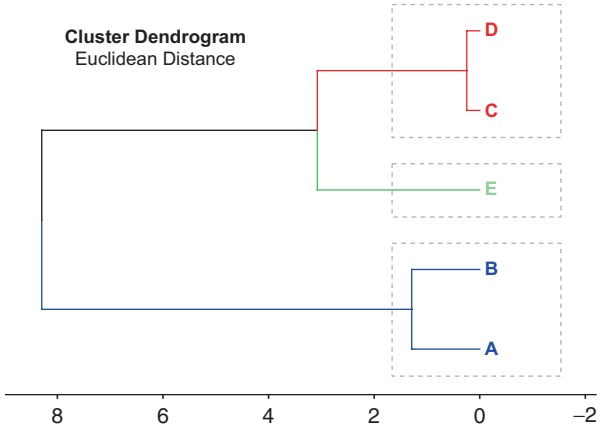

Fig. 5.6 Cluster dendrogram showing the grouping of treatments according to the characteristics of each group. Groups of different colors differ by the Euclidean distance applied in the analyzed characteristics of each treatment (see Sect. 5.5). Package: cluster (Maechler et al. 2018), factoextra (Kassambara and Mundt 2017); Functions: dist, hclust, fviz_dend

valuable contributions in magazines of high impact and potential. Therefore, the application of advanced statistical approaches is necessary and suggested so that researchers can better evaluate and present the results of different studies related to the agricultural system.

References

Dormann CF, Gruber B, Fründ J (2008) Introducing the bipartite package: analysing ecological networks. R News 8:8–11. https://doi.org/10.1159/000265935

Dunn CP, Stearns F (1987) Relationship of vegetation layers to soils in southeastern Wisconsin forested wetlands. The American Midland Naturalist 118:366–74.

Gu Z, Gu L, Eils R et al (2014) *circlize* implements and enhances circular visualization in R. Bioinformatics 30:2811–2812. https://doi.org/10.1093/bioinformatics/btu393

Kassambara A, Mundt F (2017) factoextra: extract and visualize the results of multivariate data analyses. R Packag. version 1.0.5.999

Maechler M, Rousseeuw P, Struyf A, et al (2018) Cluster: cluster analysis basics and extensions. R Packag. version 2.0.7-1

Peterson BG, Carl P (2018) PerformanceAnalytics: econometric tools for performance and risk analysis. https://cran.r-project.org/package=PerformanceAnalytics. Accessed 18 Oct 2018

Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. Ann Math Stat 27:832. 27:832. https://doi.org/10.1214/aos/1176348654

Tang Y, Horikoshi M, Li W (2016) Ggfortify: unified interface to visualize statistical result of popular R packages. R J 82:478–489

Wei T, Simko V (2017) R package "corrplot": visualization of a correlation matrix (Version 0.84)

Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer-Verlag, New York