

GUITest v 1.0: INTERFAZ GRÁFICA EN LENGUAJE R PARA TEST ESTADÍSTICOS

Osmany Molina Concepción*, Carmen C. Pons Pérez y Yaselis Guillén López.

Instituto de Investigaciones de Viandas Tropicales (INIVIT), Apartado 6, Santo Domingo, CP: 53 000, Villa Clara, Cuba.

*Autor para la correspondencia: taxonumeric@inivit.cu

Recibido: 30 de abril de 2020; Aceptado: 19 de junio de 2020

RESUMEN

Es innegable que el análisis estadístico con aplicaciones informáticas presenta ventajas para el usuario, donde este logra simplicidad de ejecución y una rápida reproducibilidad, así como se proporciona una amplia facilidad y flexibilidad en el manejo de los gráficos. Con esta motivación se creó una Interfaz gráfica para usuarios en lenguaje R que facilita realizar *tests* estadísticos en bancos de germoplasma con vistas a ayudar a los curadores en la interpretación y extracción de información útil a partir de los datos disponibles. Para esto se utilizó el R que es un programa estadístico y un lenguaje de programación de uso libre, de distribución gratuita y de código abierto. El trabajo desarrollado permite a los usuarios no técnicos aprender y realizar análisis sin necesidad de conocer el lenguaje de programación de R.

Palabras clave: aplicaciones informáticas, análisis estadístico, *software*

GUITest V 1.0: GRAPHIC INTERFACE IN R LANGUAGE FOR STATISTICAL TESTS

ABSTRACT

It is undeniable that the statistical analysis with computer applications has advantages for the user, where this one achieves a simpler and faster performance, as well as a wider facility and flexibility are given while managing graphics. That is the reason why, a graphic interface was created for R language users, which makes possible to do statistical tests in germplasm banks, so as to help the genetic users interpret and extract useful information from available data. To do so, the R language was used that is a statistical program and a programming language of free use, with a free distribution, and an open code. The work carried out allows non-technical users to learn and make analyses without knowing the R programming language.

Keywords: computer applications, statistical analysis, software

INTRODUCCIÓN

Aunque la interacción de la estadística matemática y la biología no es algo nuevo, el extraordinario desarrollo de la informática en los últimos tiempos ha contribuido a popularizar el empleo de métodos de análisis e interpretación. Un ejemplo de lo anterior lo constituye la estadística descriptiva.

Los análisis estadísticos ponen a disposición un conjunto de herramientas muy amplio, muchas de las cuales han sido desarrolladas desde hace algún tiempo, pero que han experimentado en los últimos años una importante revolución motivados por la utilización de programas computacionales como herramientas de cálculo.

Los datos se pueden analizar mediante el empleo de métodos simples o complejos, que van desde el uso de gráficos y estadísticos de tendencia central y dispersión hasta los multivariados.

Existen una gran variedad de programas estadísticos propietarios utilizados frecuentemente en el análisis y obtención de resultados, pero el empleo eficiente de estos requiere de la integración a un alto nivel de conocimientos estadísticos e informáticos para manipular correctamente sus potencialidades e interpretar adecuadamente los resultados obtenidos.

En los últimos años, ha surgido con fuerza como alternativa de *software* libre en muy distintos ambientes docentes y de investigación, un lenguaje de programación especialmente indicado para el análisis estadístico, denominado R (Ihaka y Gentleman, 1996).

R es un lenguaje de programación principalmente orientado al análisis estadístico y visualización de información cuantitativa y cualitativa. Fue oficialmente presentado en 1997 y es un *software* libre que se rige por la licencia general pública ("*General Public License*" o GPL) de la fundación de *software* libre ("*Free Software Foundation*" o GNU). Una de sus grandes fortalezas es que puede ser ampliado mediante paquetes que extienden sus funcionalidades.

El R (R Core Team, 2019) es un lenguaje de programación principalmente orientado al análisis estadístico y visualización de información cuantitativa y cualitativa. A diferencia de la mayoría de los programas que tienen interfaces tipo ventana, R es manejado a través de una consola en la que se introduce comandos propios de su lenguaje para acceder a todos los procedimientos, sin embargo, es posible diseñar interfaces gráficas para facilitar la interacción con los usuarios no familiarizados con su consola (Salas, 2008).

Por lo cual, el objetivo de esta tarea fue crear una interfaz gráfica para usuarios, en lenguaje R, que facilite realizar *test* estadísticos con vista a ayudar a los investigadores en la interpretación y extracción de información útil a partir de los datos disponibles, lo cual estimula a los usuarios no técnicos para aprender y realizar análisis sin necesidad de conocer el lenguaje de programación de R.

MATERIALES Y MÉTODOS

El trabajo se desarrolló por el grupo de Bioinformática del Instituto de Investigaciones de Viandas Tropicales (Inivit), partiendo de un diseño logrado sobre el análisis de los requerimientos expuestos por investigadores.

Se diseñaron los módulos para la gestión de archivos, edición de datos, análisis, representación gráfica y ayuda.

La programación se realizó en el lenguaje R, utilizando sus amplias potencialidades al tener disponible en su sitio en *Internet* una gran cantidad de paquetes con una vasta variedad de funciones que se pueden utilizar en la programación. Entre estos paquetes está el *tcltk* y *tcltk2* (Grosjean, 2017a, 2017b, 2017c) que permiten crear interfaces gráficas, que incluyen múltiples elementos utilizados en la confección de la interfaz de usuario, entre ellos: ventanas, menús, barras de desplazamiento, botones, botones radiales, cuadros de texto, etiquetas de texto, cuadros de selección, cajas de listas, conjuntos, entre otros.

El paquete *RODBC* (Ripley and Lapsley, 2011) facilita la conectividad con bases de datos tipo *Open Database Connectivity* (ODBC), contiene funciones que permiten abrir

una conexión con los datos que están, por ejemplo, en una hoja de *Excel* y cargarlos en R.

Además, fueron incluidas funciones pertenecientes a otros paquetes tales como: *agricolae* (Mendiburu, 2017), *corrplot* (Wei, 2017; Murdoch, 1996) y el paquete *stats* que forma parte de la librería básica de R que se instala por defecto.

Para la ejecución de la interfaz debe instalarse primeramente la consola de R y a partir de ella, teclear la instrucción (`source(GUIADtest.r)`), se llama al fichero principal que contiene la ventana de entrada. Se requiere de una plataforma *Windows®* 7 o superior y capacidad en disco duro de 750 Mb como mínimo.

RESULTADOS Y DISCUSIÓN

Como resultado se obtuvo una interfaz con un diseño atractivo, amistoso y de fácil manipulación.

La interfaz está compuesta por una ventana principal con una barra de menú en la parte superior. La primera opción de la barra de menú “Archivo” permite abrir el fichero de datos en formatos de fichero con extensiones csv o txt, además tiene las opciones de guardar área de trabajo, cambiar directorio de trabajo y salir de la aplicación (Figura 1, 2 y 3).

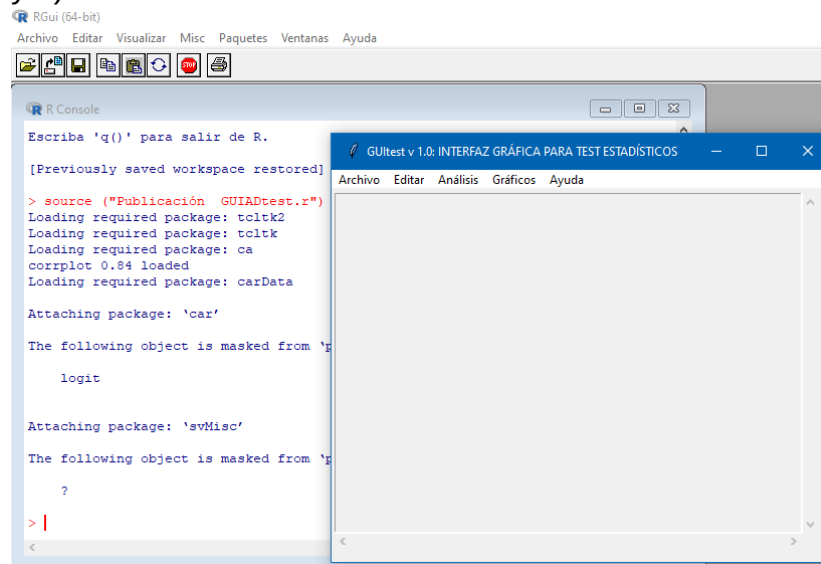


Figura 1. Interfaz gráfica de usuario para *test* estadísticos.

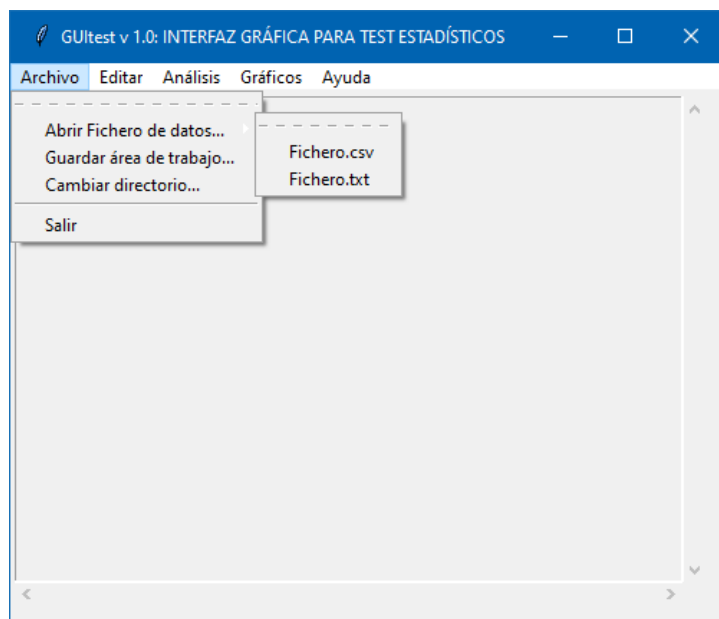


Figura 2. Opciones del menú “Archivo” de la interfaz para *test* estadísticos.

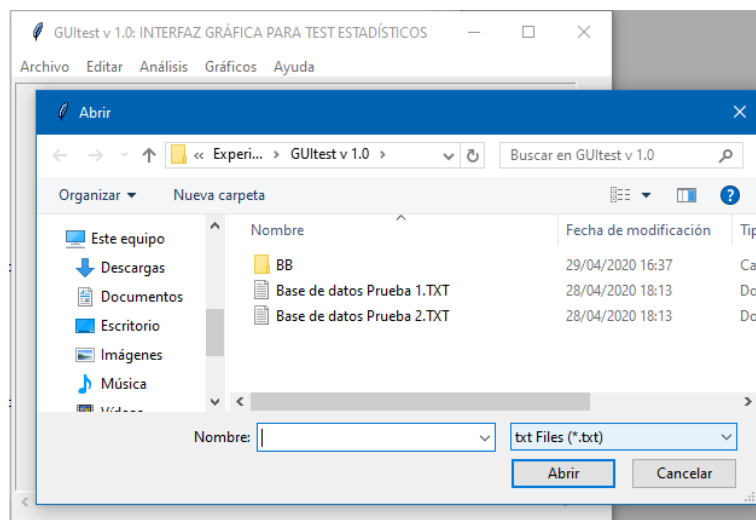


Figura 3. Opciones del menú “Archivo” de la interfaz para *test* estadísticos.

Al ejecutar “Seleccionar datos” también se imprime en la consola de R el conjunto de datos seleccionados (Tabla 1).

Tabla 1. Impresión en la consola de R del conjunto de datos seleccionados para el análisis (txt_data).

```
The imported data are in txt_data
PR Nmanos ND NDSmano Perímetro HSanas
1 22.3 7.8 112 15.3 12.4 5.590062
2 13.5 8.7 126 16.2 12.2 10.526316
3 14.7 9.2 138 16.9 13.8 24.897959
4 24.5 6.8 96 13.8 19.3 42.264151
5 13.2 7.5 122 15.9 11.8 4.918033
6 13.1 7.4 119 15.8 11.6 3.125000
7 14.5 9.2 148 17.9 11.5 0.000000
```

A través de la segunda opción de la barra de menú “Editar” se pueden “Editar datos” si fuese necesario por parte del usuario; y “Seleccionar datos” es una opción que permite seleccionar los datos para el análisis los cuales aparecerán con el nombre “txt_data”. Después de seleccionados se renombran automáticamente con “Datos.act” (Figura 4).

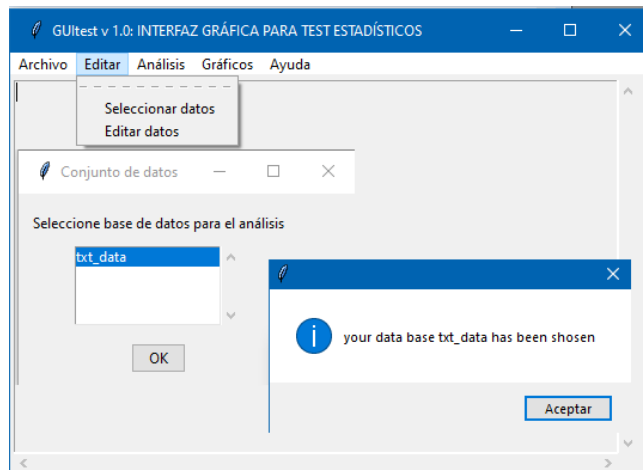


Figura 4. Opciones del menú “Editar” de la interfaz para *test* estadísticos.

Al ejecutar “Seleccionar datos” también se imprime en la consola de R el conjunto de datos seleccionado.

Mediante la opción “Análisis” (Figura 5) de la barra de menús se accede a “Resumen de datos activos” con la función *summary()* descrita en el paquete *base*, da los cuartiles de los datos cuantitativos (Tabla 2), “Test de correlación” con la función *correlation()* y “Test de Kruskal Wallis” con la función *kruskal()* descritas en el paquete *agricolae*.

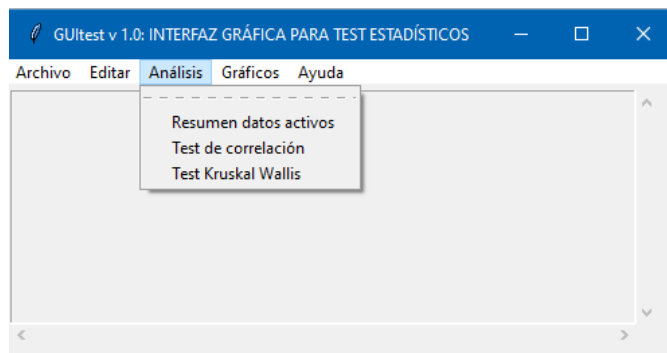


Figura 5. Opciones del menú “Análisis” de la interfaz para *test* estadísticos.

Tabla 2. Impresión en la consola de R de los cuartiles de los datos cuantitativos.

NB	Altura	AF
Min. :0.250	Min. :0.520	Min. :0.0000
1st Qu.:0.930	1st Qu.:1.173	1st Qu.:0.0100
Median :1.525	Median :1.220	Median :0.0500
Mean :1.453	Mean :1.364	Mean :0.3496
3rd Qu.:1.550	3rd Qu.:1.740	3rd Qu.:0.5500
Max. :3.010	Max. :2.160	Max. :1.1600

La segunda opción permite calcular la correlación (pearson, kendall o spearman) de las variables seleccionadas (Figura 6), además se imprimen tres tablas con la correlación de las variables, el *p-value* de estas correlaciones y el coeficiente de determinación (r^2) (Tabla 3).

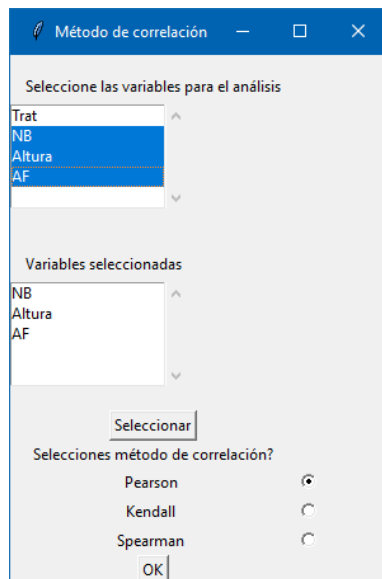


Figura 6. Opciones del menú “Análisis”: Test de correlación.

Tabla 3. Impresión en la consola de R de la correlación de las variables, el *p-value* y el coeficiente de determinación (r^2) de estas correlaciones.

```
[1] "pearson"
      NB  Altura  AF
NB      1.00   0.94 0.90
Altura  0.94   1.00 0.89
AF      0.90   0.89 1.00
[2] " p-value "
      NB  Altura  AF
NB      1      0   0
Altura  0      1   0
AF      0      0   1
[3] " r2"
      NB  Altura  AF
NB     100.00  88.36 81.00
Altura  88.36 100.00 79.21
AF      81.00  79.21 100.00
```

El tercer análisis disponible en el menú “Análisis” es el *test* de *Kruskal-Wallis* (Figura 7). El usuario debe de introducir las variables *Response* y *Group* del grupo de datos activos. Además, debe activar la casilla “Grupo SI o NO” si desea la representación entre los grupos con sus letras.

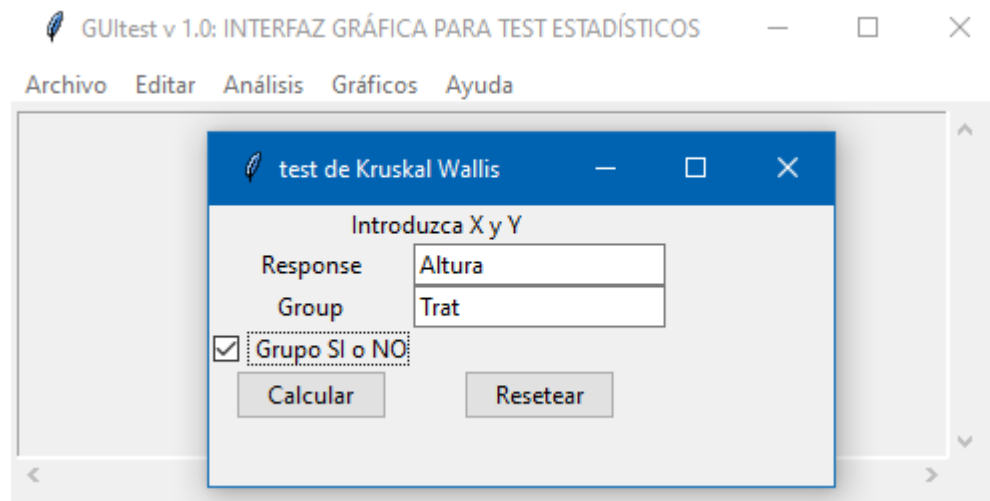


Figura 7. Opciones del menú “Análisis”: Test de Kruskal Wallis.

A continuación, se muestra el resultado de un test de Kruskal-Wallis:

```
### Command executed via Tk ###
with(Datas.act, kruskal(Altura, Trat, group = TRUE, console = TRUE,
  main = ""))
### -----
```

Study:
Kruskal-wallis test's
Ties or no Ties

Critical value: 47.17965
Degrees of freedom: 4
Pvalue chisq : 1.399023e-09

Trat, means of the ranks

	Altura	r
0	45.50	10
10	35.50	10
20	25.45	10
30	15.55	10
40	5.50	10

Post Hoc Analysis
t-Student: 2.014103
Alpha : 0.05
Minimum Significant Difference: 2.636177

Treatments with the same letter are not significantly different.

	Altura	groups
0	45.50	a
10	35.50	b
20	25.45	c
30	15.55	d
40	5.50	e

La opción del menú “Gráfico” (Figura 8) permite mostrar el gráfico de caja para las variables cuantitativas que intervienen en el análisis de correlación (Figura 9).

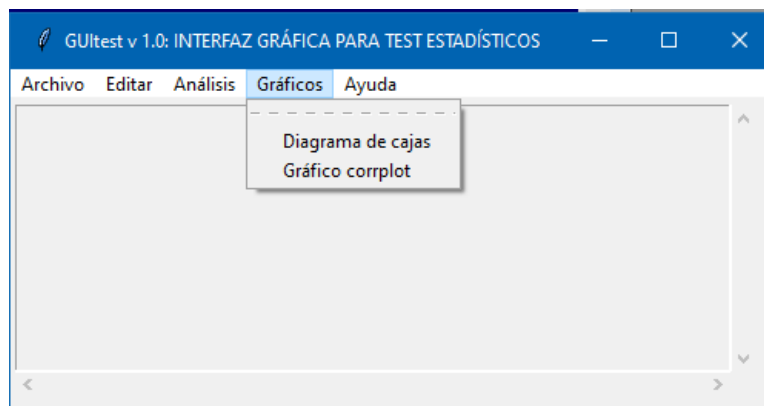


Figura 8. Opciones del menú “Gráficos” de la interfaz para *test* estadísticos.

Diagrama de caja

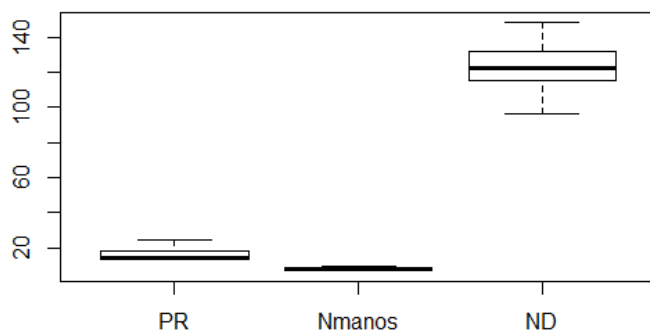


Figura 9. Diagrama de caja.

Además, la opción “Gráficos corrplot” permite mostrar los gráficos que genera la función *scatterplotMatrixcorrplot()* (Figura 10) del paquete *car* (Fox, 2019) y los gráficos obtenidos por la función *corrplot()* del paquete *corrplot* (Figura 11 y 12).

Scatter plot matrix

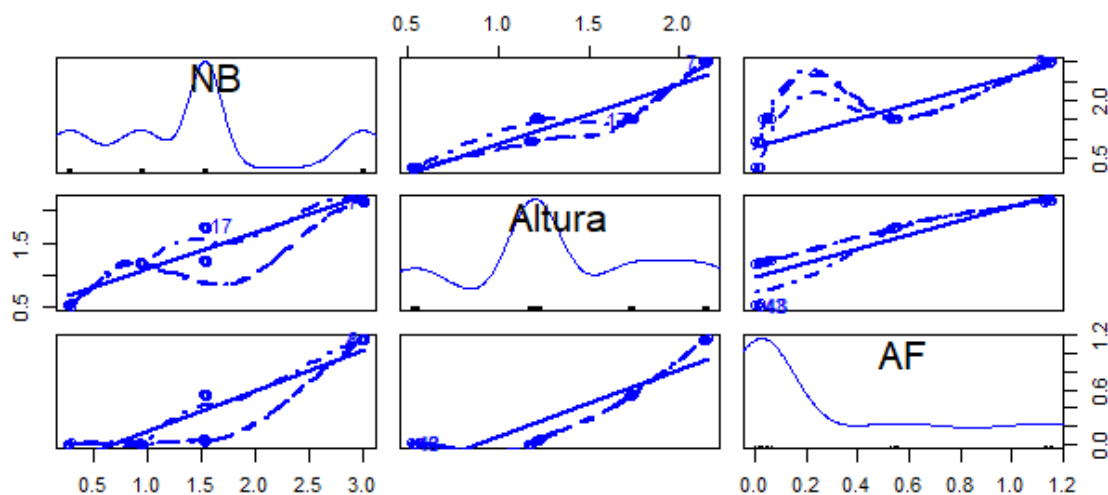


Figura 10. Gráfico de *Scatter plot* de la interfaz para *test* estadísticos.

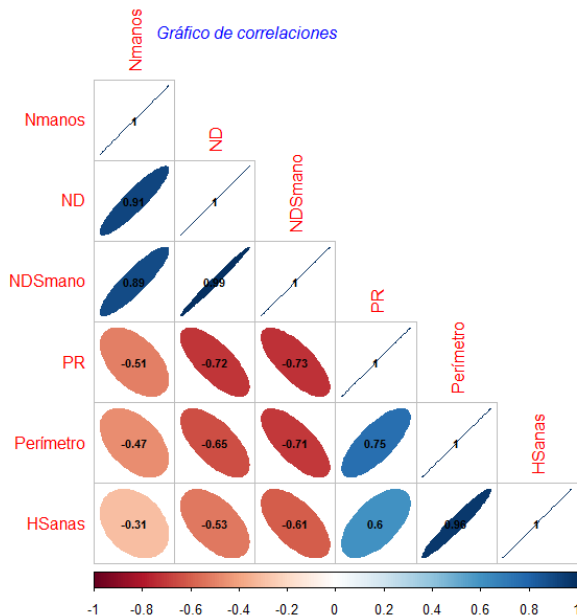


Figura 11. Gráfico de correlación entre las variables cuantitativas.

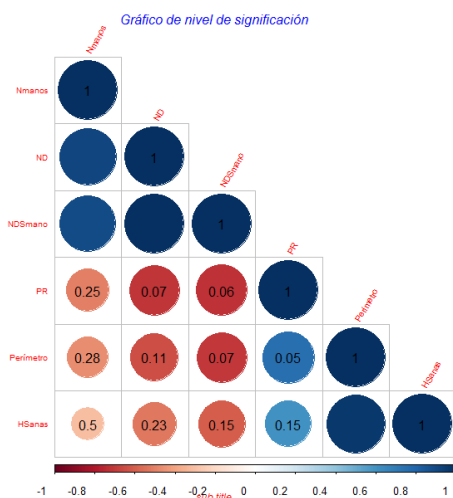


Figura 12. Gráfico de nivel de significación de la correlación entre variables.

CONCLUSIONES

Se obtuvo una aplicación informática que permite la obtención y manipulación de datos, y la ejecución de *test* estadísticos sin necesidad de conocer el lenguaje de programación de R.

La aplicación informática es un *software* libre, lo cual amplía grandemente sus posibilidades de acceso y distribución.

BIBLIOGRAFÍA

- FOX, J. and S. WEISBERG. 2019. An R Companion to Applied Regression, Third Edition, Sage.
- GROSJEAN, P. 2017. A GUI API for R. Belgium: UMONS. Consultado: 12 de abril 2017, Disponible en: <https://cran.r-project.org/web/packages/tcltk2/index.html>.

- GROSJEAN, P. 2017b. Package "tcltk2". Consultado: 27 enero de 2020, Disponible en: <http://www.r-project.org/>.
- GROSJEAN, P. 2017c. SciViews: A GUI API for R. UMONS, Mons, Belgium. Consultado: 12 de abril 2017. Disponible en: <http://www.sciviews.org/SciViews-R>.
- IHAKA, R. and R. GENTLEMAN. 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5: 299–314.
- MURDOCH, D.J. and E.D. CHOW. 1996. A graphical display of large correlation matrices. *The American Statistician*, 50:178-180.
- MENDIBURU, F. agricolae: Statistical Procedures for Agricultural Research. R package version 1.2-8. Consultado: 12 de abril 2017. Disponible en: <https://CRAN.R-project.org/package=agricolae>
- RIPLEY, B. and M. LAPSLEY. 2017. RODBC: ODBC Database Access. Consultado: 23 de mayo 2017, Disponible en: <https://cran.r-project.org/web/packages/RODBC/RODBC.pdf>.
- R DEVELOPMENT CORE TEAM. 2020. R: A Language and Environment for Statistical Computing (version 3.6.1). R Foundation for Statistical Computing. Vienna, Austria. Consultado: 27 enero de 2020, Disponible en: <http://www.r-project.org/>.
- SALAS, C. 2008. ¿Por qué comprar un programa estadístico si existe R? *Ecología Austral*, 18:223-231.
- WEI, T. and V. SIMKO. R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Consultado: 23 de mayo 2017, Disponible en: <https://github.com/taiyun/corrplot>.