Programa estadístico R, Herramienta clave en el análisis y visualización de datos

Ramírez-Valverde, Gustavo[1]; **Ramírez-Valverde, Benito**[2*]

[1] Colegio de Postgraduados Campus Montecillo. Carretera México-Texcoco km 36.5, Montecillo, Texcoco, Estado de México, México. C. P. 56230.

[2] Colegio de Postgraduados, Campus Puebla. Boulevard Forjadores de Puebla, No. 205. Momoxpan municipio de San Pedro Cholula, Puebla. C.P. 72760.

* Autor para correspondencia: bramirez@colpos.mx

Problema

El manejo, análisis y visualización de datos es parte esencial en la toma de decisiones en un ámbito científico o comercial. En la actualidad existen una gran cantidad de paquetes de programas (*software*) comerciales que realizan esta función, por ejemplo: el SAS (*Statistical Analysis Systems*), el SPSS (*Statistical Package for the Social Sciences*), y el MINITAB, entre otros, sin embargo, las licencias para su uso resultan costosas para los presupuestos que se manejan en el círculo académico, y más en tiempos de crisis económica, donde es determinante el manejo racional y adecuado de recursos económicos. El lenguaje de programación R surge bajo la filosofía "*Open Source*" y es mantenido por el "*R Development Core Team*" con la participación de una extensa comunidad de usuarios y programadores en el desarrollo de nuevas funciones, paquetes y actualizaciones que son puestas a disposición de todo el mundo en forma libre y gratuita, por lo que representa una excelente opción para el manejo, análisis y visualización de datos.

El paquete estadístico R es un lenguaje de programación orientado a objetos y su utilización requiere de programación realizada sobre una consola de comandos en lugar de una interfaz gráfica. Esto conlleva a que usar R requiere mayor esfuerzo para dominar la rigurosa sintaxis que sostiene. Este hecho se ve acentuado en personas con menor conocimiento en programación y computo, por lo que fácilmente desisten en su uso. Otro problema es que debido al grado de especialización, se presenta una subutilización del software.

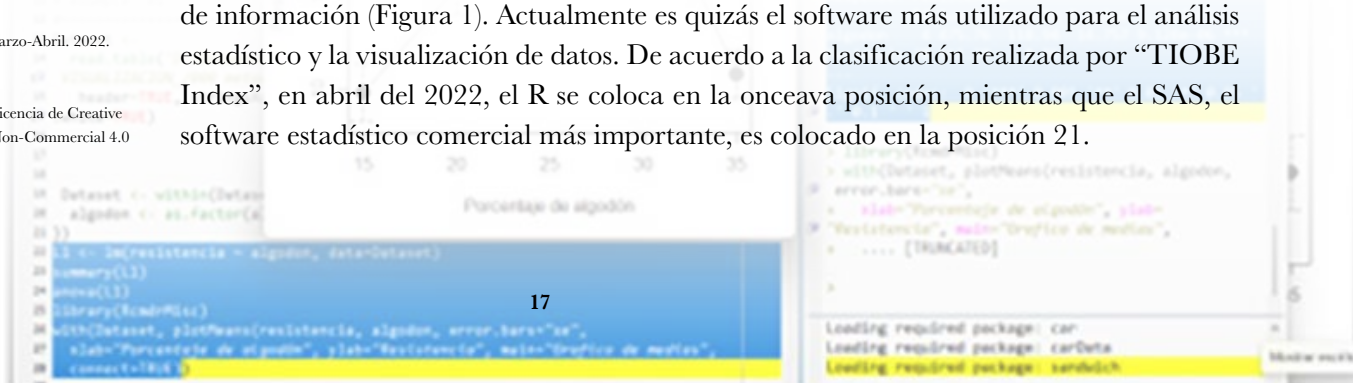
Solución planteada

El software R es un lenguaje para el análisis estadístico y gráfico, creado por estadísticos para todo el público que requieran un análisis estadístico, o lograr una mejor visualización de información (Figura 1). Actualmente es quizás el software más utilizado para el análisis estadístico y la visualización de datos. De acuerdo a la clasificación realizada por "TIOBE Index", en abril del 2022, el R se coloca en la onceava posición, mientras que el SAS, el software estadístico comercial más importante, es colocado en la posición 21.

Cómo citar: Ramírez-Valverde, G., & Ramírez-Valverde, B. (2022). Programa estadístico R, Herramienta clave en el análisis y visualización de datos. *Agro-Divulgación*, 2(2).

Editores académicos: Dra. Ma. de Lourdes C. Arévalo Galarza y Dr. Jorge Cadena Iñiguez.

Agro-Divulgación, 2 (2). Marzo-Abril. 2022. pp: 17-22.



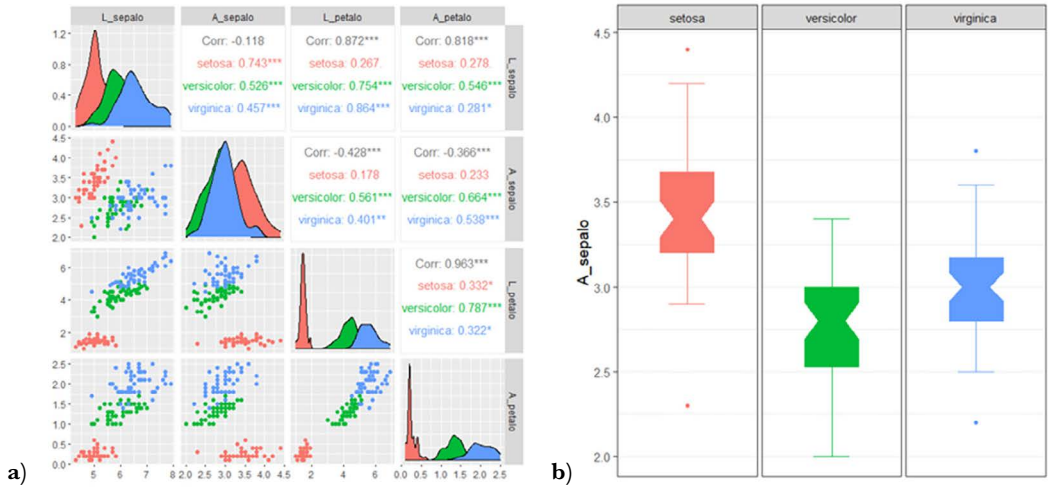


Figura 1. Ejemplos de graficos en R: a) Matriz de diagramas de dispersion con informacion adicional; y b) Gráfico de cajas por especies

Los derechos de autor del código fuente principal de R pertenecen a la "*Fundación R*" y se publican bajo la Licencia Pública General (GNU), que permite:

1. Ejecutar el programa, para cualquier propósito.
2. Estudiar cómo funciona el programa y adaptarlo a sus necesidades.
3. Redistribuir copias.
4. Mejorar el programa y divulgar sus mejoras al público, de modo que toda la comunidad se beneficie.

Los orígenes de R se remontan a 1976 cuando Richard A. Becker, John M. Chambers y Allan R. Wilks inventaron el lenguaje S (antecesor del R) en los laboratorios AT&T Bell (USA), programa que evolucionaria por un lado en el software comercial S-Plus, que tuvo una gran aceptación y por otro lado el R (Ross Ihaka y Robert Gentleman, 1990) como una versión gratuita; el S-Plus y el R tienen una gran compatibilidad, con la diferencia de que el S-Plus representó un software mucho más amigable. En 1995, se crea el GNU "*General Public License*" para hacer libre al R y en 1997 se crea el "*R Core Team*" que es el grupo que actualiza y da mantenimiento al proyecto: "*R: The R Project for Statistical Computing*".

Algunos aspectos que ponen de manifiesto las ventajas de R son:

1. Paquete con la filosofía "*Open source*". Lo que lo hace un paquete de acceso universal y gratuito.
2. Amplias posibilidades y capacidad de manejo de datos.
3. Es un proyecto flexible que permite actualizar rápidamente las técnicas avanzadas que la comunidad científica genera.
4. Permite la creación de gráficos de alta calidad.
5. Los gráficos generados son fácilmente exportables a otros formatos: PostScript, pdf, bitmap, pictex, png, jpeg, etc.

6. Se mantiene y actualiza por la contribución voluntaria de la comunidad científica (se actualiza al menos dos veces al año).
7. Permite a los usuarios crear sus propias funciones.
8. Los paquetes (conjunto de rutinas con funciones específicas) generados por la comunidad científica (actualmente 19,022 paquetes) están disponibles y en constante actualización.
9. Consume pocos recursos informáticos.
10. Está disponible para todos los sistemas operativos (Windows, Macintosh y sistemas Unix).
11. R puede interactuar con otros paquetes como SPSS, SAS, Excel, etc.
12. Puedes crear informes reproductibles y en varios formatos (pdf, word, html).
13. Permanentemente se está mejorando la facilidad de manejo del software.

La gran desventaja del R lo representa el hecho de que se tenga que programar en una consola de comandos, sin embargo, el software presenta herramientas y utilidades para poder desarrollarlas, actualmente existen diversos intentos por diseñar una interfaz que convierta al R en un software amigable. El desarrollo de estas interfaces se podría agrupar en dos vertientes:

I) Soluciones basadas en un entorno integrado de desarrollo (IDE, por sus siglas en inglés). Este tipo de solución incentiva el uso de R en usuarios principiantes y facilita el aprendizaje del lenguaje (Figura 2). En este enfoque podemos destacar:

- **RStudio**. Es IDE de código abierto para R que permite interactuar de manera muy simple, incluye una consola, un editor de resaltado de sintaxis que admite la ejecución directa de código, así como herramientas que facilitan el uso de R (también es compatible con Python) y está disponible en forma gratuita para varias plataformas (Windows, Mac y Linux). Es quizás la forma más popular de uso del software R y cuenta con una versión comercial que extiende su potencialidad. RStudio pueden ser descargadas desde: http://www.rstudio.com.
- **RKWard**. Es un desarrollo integrado (IDE), aunque pretende también llegar a ser una "interfaz gráfica de usuario" (GUI, véase próxima sección para una mejor descripción de GUI), y proporciona una excelente herramienta para administrar diferentes tipos de objetos de datos; incluso permitiendo la edición perfecta de ciertos tipos. El objetivo de RKWard es "proporcionar una interfaz R portátil y extensible para aplicaciones básicas y análisis estadístico y gráfico avanzado, sin comprometer la flexibilidad y modularidad del propio entorno de programación R". Su proceso de instalación es un poco más complicado y tardado; se puede obtener de: http://rkward.kde.org.
- **Tinn-R**. Es un editor/procesador de textos ASCII/UNICODE genérico para el sistema operativo Windows, muy bien integrado en R, con características de Entorno de Desarrollo Integrado (IDE) aunque también presenta características de Interfaz Gráfica de Usuario (GUI). Se descarga desde https://tinn-r.org.

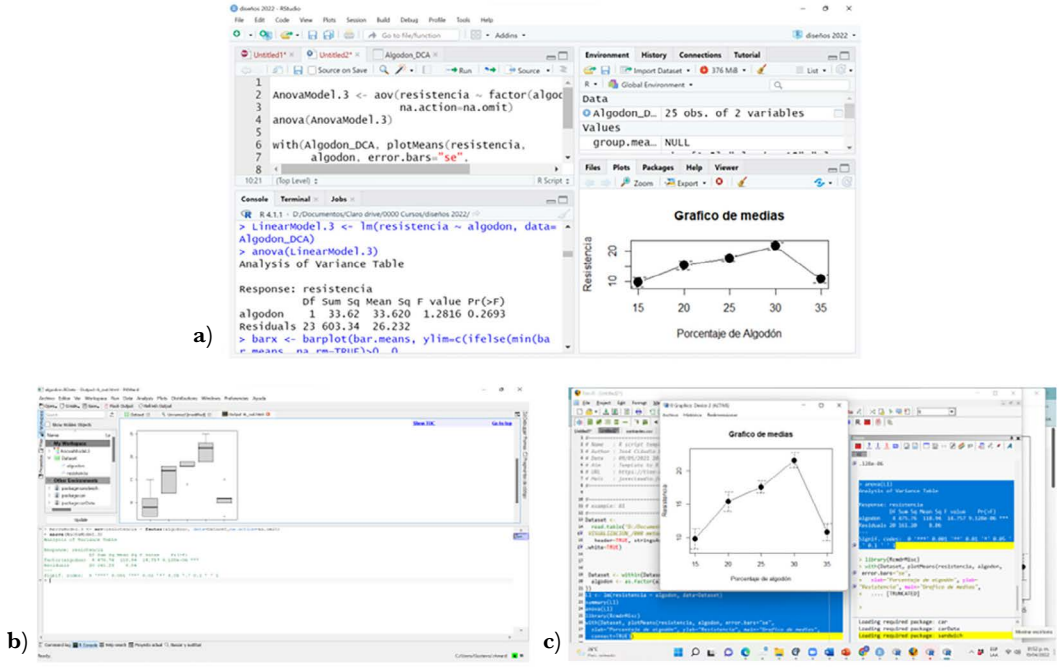


Figura 2. Pantallas muestra de diferentes IDE para el manejo simple de R: a) RStudio; b) RKWward; y c) Tinn-R.

II) **Soluciones basadas en una plataforma avanzada de interfaz gráfica de usuario** (**GUI**). Este tipo de soluciones priorizan el ambiente gráfico, logrando soluciones totalmente amigables, que incluso pueden ser usadas aun cuando se desconozca el lenguaje R y su sintaxis (Figura 3). En este enfoque se puede destacar:

- **Blue-Sky**. Esta GUI de R puede ser descargada gratuitamente y proporciona una interfaz que puede ser usado sin conocer R (existe una versión comercial con mayor potencialidad), permite obtener el código de R que genera los resultados, a la vez que permite modificarlo para aumentar su potencialidad para lograr resultados no disponibles automáticamente (tener toda la potencialidad de R), además despliega un acceso excelente a las ayudas de R. Presenta un énfasis en aplicaciones de minería de datos e inteligencia artificial. Sin embargo, resulta en un software un poco pesado, pero con un excelente manejo de archivos muy similar al del software comercial SPSS.
- **R-UCA**. El proyecto R-UCA construye un sistema totalmente funcional para Windows que se basa es una recopilación de R junto a R-Commander y a algunos paquetes de uso frecuente, no se requiere conocer la sintaxis de R, pero no incluye todas las funciones que contiene R, sin embargo, al realizar un procedimiento, muestra al mismo tiempo el código en R que genera los resultados, permitiendo modificarlo o ampliarlo a las funciones no incluidas en el R-Commander. Tiene la desventaja de que solo esta implementado para la plataforma Windows para 32

bits; además de que tiene un desfase (para asegurar 100% de funcionalidad) con la versión de R, esto es, la última versión funcional es con una versión anterior de R. Tiene la ventaja de que se instala en un único paso R, R-Commander y los otros paquetes recomendados, además de que permite instalar R en una computadora sin conexión a internet (R-UCA: http://knuth.uca.es/version_R-UCA.php).

- **Jasp**. Es un proyecto de código abierto apoyado por la Universidad de Ámsterdam, muy amigable con una interfaz intuitiva diseñada pensando en el usuario, con gran semejanza al software comercial SPSS, cuya operación aparece independiente del software R, aunque todas las operaciones son realizadas en R (aun cuando no lo hace evidente). Tiene el problema que es muy difícil obtener el código de R que genera los resultados y por consiguiente modificarlo, aunque permite generar programas en R con los datos incluidos en los programas. Otra gran ventaja es que en su ambiente gráfico incluye análisis clásico y bayesiano (Jasp: https://jasp-stats.org).
- **EZR** (**Easy R**). Es un proyecto japonés (2021) con la misma idea del R-UCA, esto es, se basa en R y "R-Commander", pero aumenta su funcionalidad con énfasis en aplicaciones médicas. Incluye todas las funciones de R-UCA, pero aumenta funciones estadísticas que se utilizan con frecuencia en estudios clínicos, como los análisis de supervivencia, el análisis de riesgos competitivos; el uso de covariables dependientes del tiempo, metaanálisis, cálculo del tamaño de la muestra, entre otros. Tiene además la ventaja de que a diferencia del R-UCA puede ser utilizado con 32 o 64 bits, además de que existe una versión para plataformas Mac (OS X), (EZR: https://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmedEN.html)
- **InfoStat**. Es un software para análisis estadístico de aplicación general desarrollado bajo la plataforma Windows de uso muy simple (con semejanzas al JMP de SAS y al MINITAB), es incluido en esta parte porque además de que incluye análisis realizados con el motor del software R, facilitando en forma gráfica la operación de modelos complejos de R (modelos mixtos, modelos generalizados mixtos y modelos no lineales mixtos), incluye una versión IDE que permite realizar análisis en R independientes de InfoStat, lo que le da gran potencialidad. El software InfoStat es comercial, desarrollado por una universidad argentina con un precio de recuperación (50 dólares por una licencia personal), pero puede ser usado con algunas limitaciones en forma gratuita (InfoStat: http://www.infostat.com.ar).
- **Jamovi**. Es un software gratuito basado en R, parte de su intencionalidad es facilitar la transición el de un software como SPSS al Jamovi y se encuentra disponible para Windows, Mac (Os x) y Linux. Tiene mayor número de funciones en cuanto a métodos clásicos, pero en cuanto a métodos bayesianos es menos versátil. Se puede obtener fácilmente el código que generan los métodos que realiza en R (Jamovi: http://www.jamovi.org).

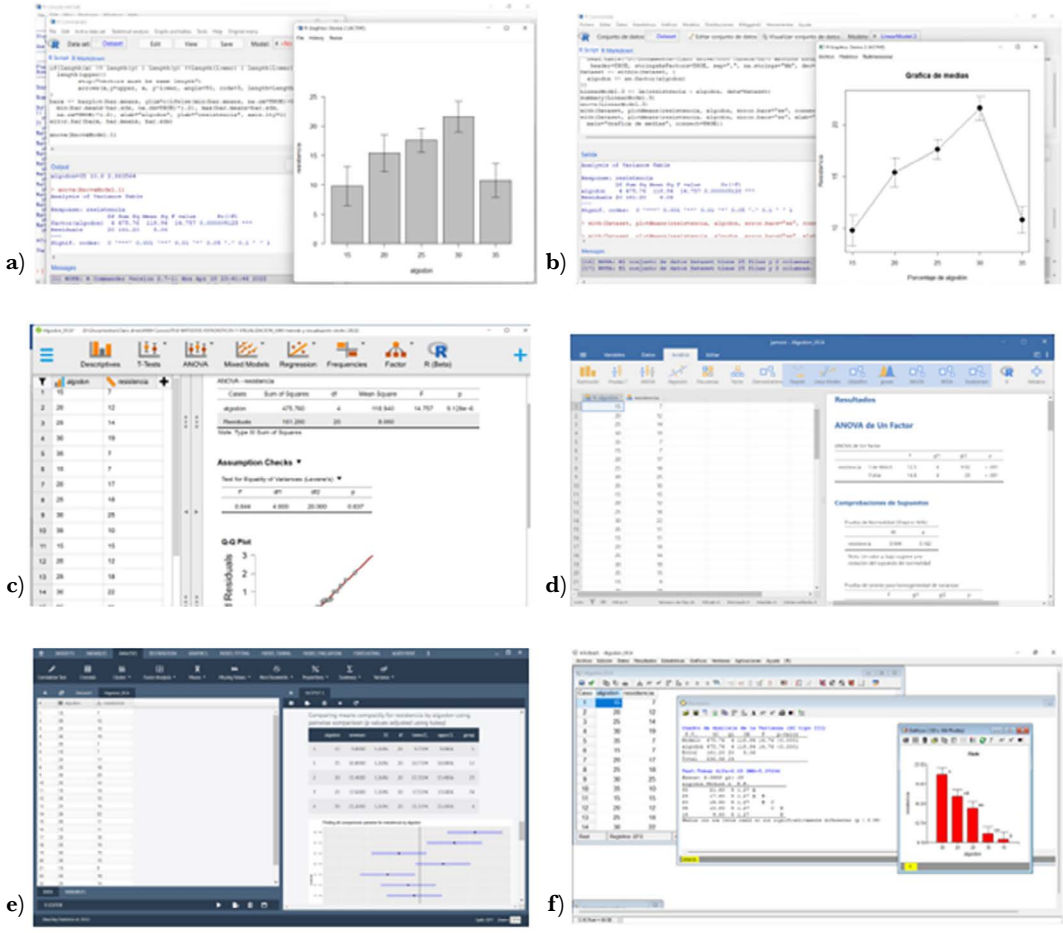


Figura 3. Pantallas muestra de diferentes GUI para el manejo simple de R: a) R-UCA; b) EZR; c) Jasp; d) Jamovi; e) BlueSky; y f)InfoStat.

El software estadístico R es totalmente apropiado para la estadística descriptiva e inferencial que requieren los estudios agronómicos y relacionados con el medio rural. Es posible realizar análisis de experimentos, análisis multivariados, análisis de datos categóricos, meta-análisis; minería de datos, redes neuronales e inteligencia artificial, entre otros.

Es importante destacar que existen varios repositorios en diversas partes del mundo donde es posible obtener de forma gratuita este software, siendo el Colegio de Postgraduados uno de estos sitios. Usted lo puede obtener libremente desde la dirección: http://www.r-project.org.