REVISTA CUBANA DE CIENCIAS BIOLÓGICAS
http://www.rccb.uh.cu

CARTAS AL EDITOR

Falta una letra en el abecedario de los ecólogos cubanos: la R

R: a missing letter in alphabet of Cuban ecologists

Dennis Denis Ávila[1*] **y Víctor Manuel Ramírez-Arrieta**[1]

[1] Departamento de Biología Animal y Humana, Facultad de Biología, Universidad de La Habana, Cuba

[2] Instituto de Ciencias del Mar, Cuba

*Autor para correspondencia: *dda@fbio.uh.cu*

Recibido: 2020-07-16

Aceptado: 2021-01-26

INTRODUCCIÓN

La Ecología a lo largo del pasado siglo sufrió una metamorfosis desde una ciencia descriptiva y naturalista hacia una rama fuertemente cuantitativa. En las últimas décadas, con el explosivo desarrollo de la informatización y las nuevas tecnologías, la producción de datos ha conducido a la que se llama la Era del Big – Data. Los sensores automáticos, los instrumentos satelitales para el monitoreo de la superficie del planeta, los equipos de secuenciación automática, la ciencia ciudadana – impulsada por el desarrollo de las redes sociales y megaproyectos internacionales de integración de información en el novedoso campo de la informática de la Biodiversidad, han llevado a que los científicos deban manejar cantidades crecientes de datos, estructurados y no estructurados. De igual forma, por la propia madurez del campo, las preguntas a responder son cada vez más sofisticadas, complejas y multidimensionales.

Enfrentar estos cambios no puede hacerse con las herramientas tradicionales (Higgs 2015; Touchon y McCoy, 2016). Por ello las habilidades de programación de códigos informáticos se han incorporado como una de las herramientas básicas en la formación de un investigador en esta rama. Como parte esencial de las investigaciones, la mayoría de los ecólogos en la actualidad suelen escribir sus propios códigos para chequear, manipular, analizar y visualizar sus datos (Mislan *et al.* 2016). Varias herramientas informáticas son apropiadas para esto (ej.: *Python*, *MATLAB* y *SAS*), pero en las últimas décadas ha despuntado el lenguaje de programación R (R *Development Core Team* 2017) como uno de los más potentes y de mayor popularidad (Bollmann *et al.,* 2017).

R es ampliamente utilizado en muchos campos científicos y técnicos (Viswam y Srinivasa, 2019). Lai *et al.* (2019) describieron las tendencias en su utilización en 30 importantes revistas de Ecología (Factor de impacto > 3,0) durante la década de 2007 – 2017, a partir de una muestra de más de 60 900 artículos. Según estos autores, el número de publicaciones que usó R como herramienta principal para el análisis de datos ascendió al 33,5% (20 395 artículos), pero ha experimentado un aumento lineal desde 11,4% en 2008 hasta 58% en 2017. En la actualidad, más de la mitad de las publicaciones en revistas de impacto de Ecología utilizan esta herramienta, lo que ha llevado a varios autores a afirmar que R es un componente significativo del análisis de datos contemporáneo.

De hecho, recientemente, una nueva disciplina con el nombre de Ciencia de Datos (*Data Science*) ha emergido en el contexto mundial, abarcando una gama de herramientas de estadística, aprendizaje de máquina y métodos computacionales en apoyo de las empresas y otras actividades que se guían por la información de grandes volúmenes de datos para orientar sus acciones. Muchas empresas que usan Ciencia de Datos están utilizando R como núcleo y persiguen ansiosamente a programadores capacitados en este lenguaje, ofreciendo empleos de los mejores pagados del mundo actual (salarios que promedian más de $117 000 al año). Campos tan diversos como las finanzas, economía, medicina o producción industrial utilizan primariamente este lenguaje de programación. Facebook utiliza R para el análisis de las redes sociales y Twitter, lo usa para el análisis semántico y las visualizaciones de sus temas. Como un indicador significativo de su importancia en la actualidad cabe mencionar que hasta existe un buscador especializado – tal como *Google Académico* o *DataSearch* – apoyado por Google y enfocado únicamente en recursos de R: RSeek (https://rseek.org).

Sin embargo, cuando se revisa el escenario de las investigaciones ecológicas en Cuba, en las publicaciones sobre biodiversidad, ecología y medio ambiente solo existen casos aislados que han utilizado este programa (ej: Fontenla *et al*., 2019, Ramírez-Arrieta y Denis, 2020). En esta comunicación, se analizan las causas y consecuencias de esta ausencia y se hace un llamado a la incorporación de esta importante herramienta en los currículos de pregrado y postgrado sobre la base de las ventajas y posibilidades que brinda.

El "abecedario"…

Antes de comenzar a hablar de la "letra ausente" a la que, metafóricamente se refiere el título de esta comunicación, ameritaría mencionar brevemente el resto del "abecedario": el universo de programas estadísticos que se utiliza en las investigaciones ecológicas cubanas. Para ello se revisaron todas las publicaciones de los últimos tres años, en cinco de las revistas científicas cubanas más productivas actualmente en este campo: la Revista del Jardín Botánico Nacional, la Revista del Centro de Investigaciones Marinas de la Universidad de La Habana, la Revista Cubana de Ciencias Biológicas, Acta Botánica Cubana y Poeyana. Entre todas se identificaron 112 artículos con resultados originales que podrían incluir la producción o uso de datos cuantitativos.

No se tuvieron en cuenta las listas de especies, descripciones locales de flora o fauna, artículos de taxonomía o sistemática, nuevos reportes, revisiones y demás tipos de publicaciones que no suelen incluir análisis numéricos extensos. De la muestra se excluyeron otros 29 trabajos que, al ser revisados, se encontró que no requerían procesamientos estadístico (trabajos de histología, de procedimientos experimentales o de laboratorio, con caracterizaciones verbales).

En la muestra restante (83 artículos) se identificaron 68 menciones a programas estadísticos, al existir 20 trabajos que, a pesar de realizar este tipo de análisis, no informaban el programa utilizado y algunos otros referían el empleo de varios programas simultáneamente. El 40% de las menciones correspondió a versiones del programa *Statistica*, de la compañía *StatSoft*, del cual se utilizaron sus versiones 6.0 a la 10.0. Le siguió, con un 25% de las referencias, el programa *Paleontological Statistics* (PAST), un programa gratuito para investigaciones paleontológicas desarrollado y mantenido por Øyvind Hammer, del museo de Historia Natural de la Universidad de Oslo, entre 1999 y 2012, en sus versiones 1.75 a 3.14 (Hammer *et al.,* 2001). Este programa tiene un buen balance entre simplicidad y potencia, lo que lo hace una buena opción para trabajos relativamente sencillos, pero es poco flexible en sus opciones y las salidas gráficas no tienen gran calidad.

Un 10% de las referencias a programas estadísticos correspondió con el programa PRIMER (*Plymouth Routines In Multivariate Ecological Research*), que fue utilizado exclusivamente en Biología marina, posiblemente por acompañar el libro de Clarke y Warwick (2001) enfocado en este campo. Del resto de los programas identificados, el *Statistical Packages for Social Sciences* (SPSS) y el *StatGraphics*, conjuntamente, aportaron el 8% de los reportes. Estos fueron seguidos de otros, de apariciones esporádicas, como *GraphPad Instat*, *InfoStat* y *SAS*.

De todos los artículos revisados, solo uno empleó el programa R (Fontenla *et al.,* 2019). Por supuesto, esto no describe la totalidad de la actividad científica en el campo de la Ecología cubana sino que se restringe a una parte de los resultados publicados a nivel nacional. Si la muestra se extendiese a publicaciones internacionales esta cantidad posiblemente aumentaría de forma discreta. Contrastando estas estadísticas con las mencionadas previamente a nivel internacional, las diferencias son evidentes.

¿Por qué R?

El ordenamiento por popularidad de los lenguajes de programación según el índice *RedMonk* (*https//redmonk.com*) en su versión de 2020 ubica a R entre los programas topes y de más rápido crecimiento (Fig. 1). Este indicador, que se publica bianual, se basa en el uso de los programas para proyectos de *GitHub* y en las discusiones sobre el lenguaje en *StackOverflow*. En relación al uso actual, R sobrepasa a todos los demás programas enfocados a datos, como Matlab y SAS.

Sin embargo, la popularidad de un programa no es un argumento sólido para proponer su uso, sino que habría que identificar y analizar las causas de dicha popularidad. R es mucho más que un programa estadístico. Es un lenguaje y ambiente de programación que, si bien surgió enfocado al análisis de datos, es muy apropiado para cualquier tipo de investigación ya que permite el desarrollo de aplicaciones que automatizan los procesos de análisis, además de cualquier otro tipo de tarea que pueda programarse. Es un lenguaje interpretado (sus comandos se ejecutan sin necesidad de un compilador), es multiplataforma (puede ejecutarse en cualquier sistema operativo) y es de código abierto, creado bajo la licencia GNU (*General Public Licence*), por lo que no hay restricciones para su uso.

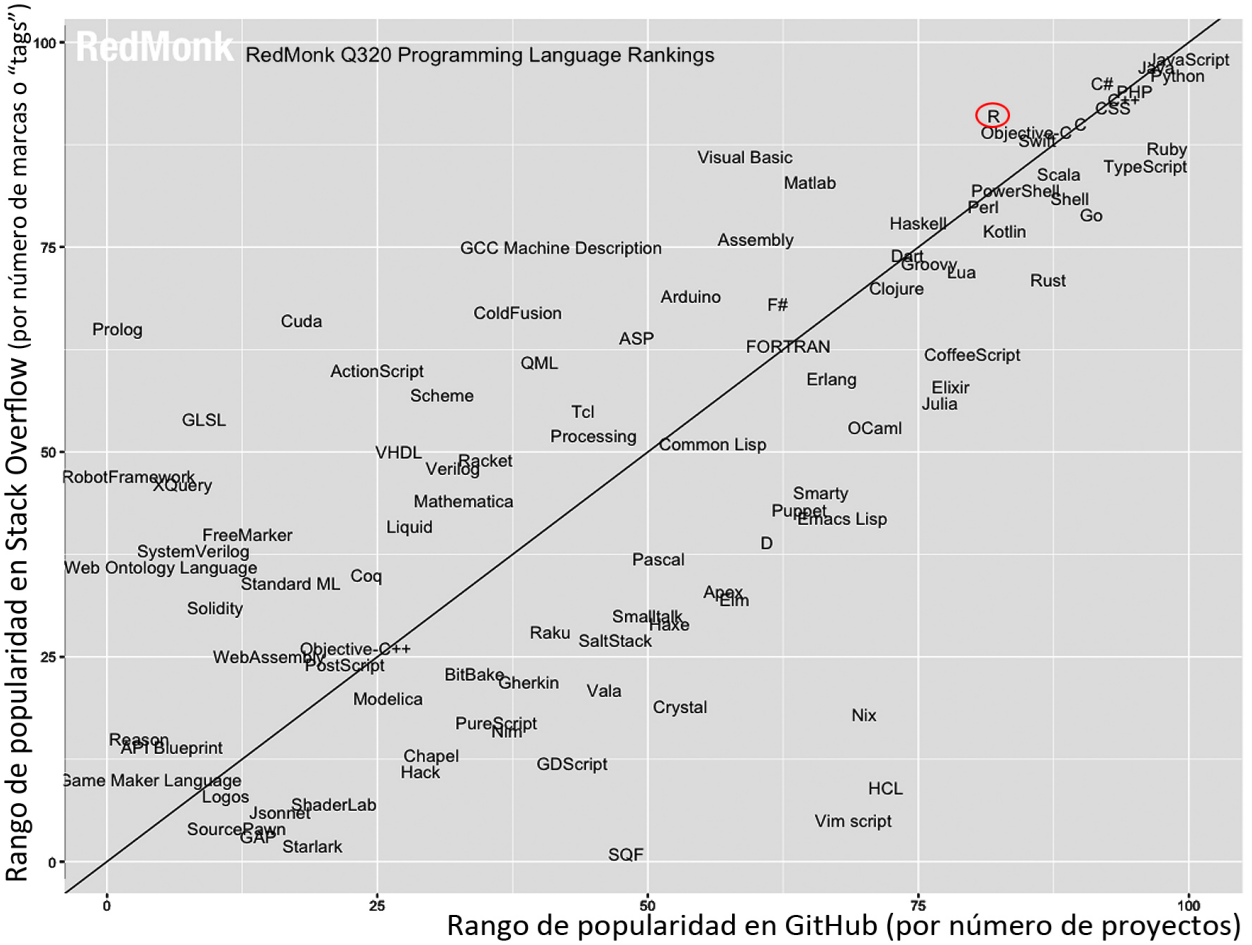

Figura 1: Ordenamiento de los lenguajes de programación según su popularidad en *GitHub* y *StackOverflow*, según el Índice *RedMonk* (tomado de *https//redmonk.com*). Se resalta en rojo la ubicación de R.

Figure 1: *Ordering of programming languages by popularity in GitHub and StackOverflow, according to RedMonk index (from https//redmonk.com). R location is highlighted in red.*

El proyecto original de R fue publicado por Robert Gentleman y Ross Ihaka, profesores del Departamento de Estadística de la Universidad de Auckland en Nueva Zelanda en 1996, como una implementación del lenguaje de programación S, específicamente enfocada en estadísticas y gráficos. Inicialmente, se restringió al mundo académico pero luego se generalizó por el mundo empresarial, con lo cual se encuentra entre los lenguajes estadísticos de más rápido crecimiento. En el mundo del reciente campo de la Ciencia de Datos, R es el lenguaje más popular y es intensamente utilizado, tanto para datos estructurados como no estructurados. Actualmente, es enseñado en la mayoría de las universidades del mundo. En Biología se ha hecho popular por el desarrollo del proyecto Bioconductor (http://www.bioconductor.org) que provee herramientas para el análisis específico de datos genómicos. La última versión de Bioconductor (3.11) contiene 1903 paquetes de códigos, 391 paquetes de datos experimentales, 961 paquetes de anotaciones y 27 flujos de trabajo automatizados.

¿Cuáles serían las ventajas más importantes del uso de R?

1 - Carácter gratuito

La primera ventaja del programa radica en que es gratuito. Incluso para el análisis de datos más básico, si bien es posible lograr los mismos resultados empleando otras herramientas como *Statistica*, *SAS*, *SPSS* o *MATLAB*, debe recordarse que estos son programas patentados y comerciales y su costo es limitante. En contextos donde las leyes de propiedad intelectual son de implementación estricta se penaliza su adquisición por métodos ilegales (versiones *crackeadas*, piratería informática). Posiblemente, sea esta la razón por la cual varios autores ya han detectado una disminución global en el uso de estos programas comerciales en estas ramas de la ciencia (Tippmann, 2015; Muenchen, 2017) (http://r4stats.com/articles/popularity/). Touchon y McCoy (2016), en una encuesta a autores de siete revistas principales de Ecología, detectaron también que el uso de *SAS* y *SPSS* había disminuido de 2006 a 2013.

2 - Estructura modular (por paquetes)

R es un entorno de programación simple y con relativamente pocas funciones en su forma básica. Al ser instalado, incluye un conjunto mínimo de paquetes (en inglés *packages*) que incrementan sus posibilidades. Los paquetes son colecciones de funciones, otros códigos y datos de muestra que permiten hacer tareas adicionales que no existen en el programa base. Esta estructura, por módulos que se seleccionan e instalan según las necesidades de cada usuario, es una fuerte ventaja que se asocia a la gran cantidad de recursos que ya existen y que se van creando continuamente, lo cual sería su tercera ventaja.

3 - La enorme cantidad de recursos que ofrece

En el sitio CRAN (*Comprehensive R Archive Network*, https://cran.r-project.org/) hay disponibles más de 13000 paquetes, con una tasa de crecimiento exponencial, que no muestra signo alguno de disminución a corto o mediano plazo (Fig. 2). Entre estos existen, al menos, 3000 funciones específicas para investigaciones ecológicas. Puede asegurarse que R ofrece en la actualidad prácticamente todos los modelos estadísticos existentes, así como numerosas rutinas para la exploración de datos. Según las estadísticas de este propio sitio (https://CRAN.R-project.org/view=Environmetrics) hay más de 100 paquetes usados con alta frecuencia en el análisis de datos ecológicos y ambientales. La frecuencia de uso de un paquete depende de muchos factores, por ejemplo, la fecha de lanzamiento, el número de funciones y su alcance. Se ha descrito que la popularidad de los paquetes puede incluso afectar las principales tendencias de los métodos estadísticos empleados para el análisis de datos ecológicos.

Lai *et al.* (2019) en su revisión describieron el uso de más de 2400 paquetes en artículos de Ecología. De ellos, 31 fueron utilizados en más de 100 investigaciones. El paquete con mayor frecuencia fue *lme4*, que se emplea para ejecutar modelos lineales generalizados (Bolker *et al.,* 2009), seguido de *vegan*, un paquete ampliamente utilizado para el análisis multivariado en ecología comunitaria así como con otras funciones basadas en matrices de disimilitud (por ejemplo, la prueba de Mantel y el Análisis de Componentes Principales), por lo que es muy utilizado también en biología molecular y ecología microbiana (Excoffier *et al.,* 1992; Dixon, 2003, Phipson y Smyth 2010; Diniz-Filho *et al.* 2013). El paquete *APE* es el cuarto paquete más popular entre los ecólogos por el aumento en el número de estudios que incorporan información filogenética en la ecología comunitaria (Swenson, 2014; Cadotte *et al.,* 2017).

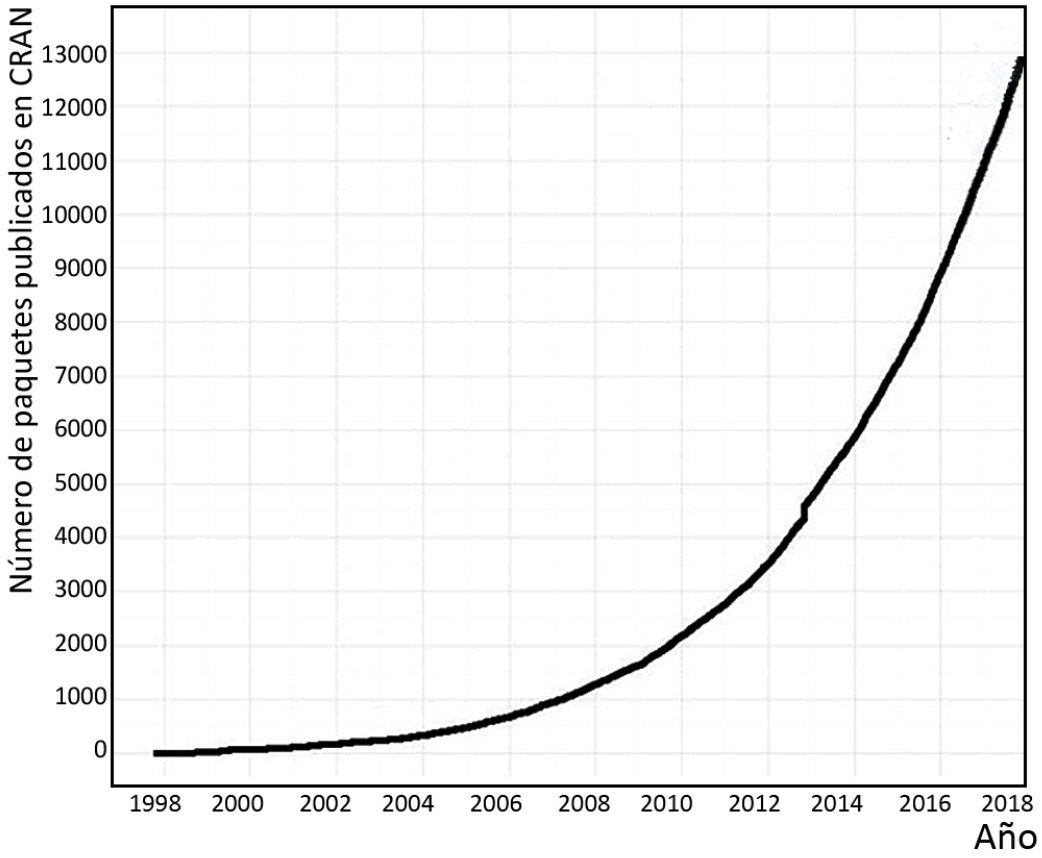

Figura 2: Comportamiento de la tasa de incorporación de nuevos paquetes en CRAN a lo largo de los años. Se incluyen en el conteo paquetes que luego han sido retirados, por ello sobrepasa el número actual de paquetes presentes.

Figure 2: *Trend in new packages uploaded to CRAN along the years. In the count packages later dismissed are included, that is why number overflow current packages total.*

4 - Los paquetes de R permiten la aplicación rápida de métodos nuevos

Como son tantos los estadísticos y ecólogos cuantitativos que usan R como su entorno de trabajo, la inmensa mayoría de los métodos estadísticos nuevos son, primero, desarrollados en R (Bivand *et al.,* 2013; Swenson, 2014; Borcard *et al.,* 2018). Por ello, los nuevos enfoques teóricos y sus métodos asociados se implementan y generalizan con gran rapidez. Los paquetes generados en estas investigaciones de borde posibilitan la aplicación instantánea de nuevas herramientas bioestadísticas y de modelado que antes eran inaccesibles para la mayoría hasta que apareciese un software que las implementara (Simpson, 2018).

5 - Los códigos de R permiten la automatización de actividades repetitivas

Con frecuencia en las investigaciones existen tareas repetitivas que consumen mucho tiempo y esfuerzo. Por ejemplo, si se obtienen múltiples ficheros de la salida automática de un equipo o se tienen distintas descargas de bases de datos generalmente se necesita filtrarlos, limpiarlos e integrarlos en una sola matriz. También, cada vez que se toman datos en una expedición, es necesario integrarlos a las matrices generales y chequear su consistencia para detectar posibles errores o discordancias, de forma mucho más rápida y que no depende de la inspección visual, tan propensa a errores humanos. En lugar de repetir manualmente una serie de operaciones una y otra vez sobre cada

fichero, simplemente se programa un código para una función que se repita. Por ejemplo, en el código disponible en el sitio https://gist.github.com/daroczig/ puede recrearse la figura 2, con datos actualizados que se descargan automáticamente de Internet.

Durante una investigación es frecuente que uno tenga que repetir los análisis y rehacer las figuras hasta cuatro o cinco veces, ya sea porque aparecen nuevos datos o se detectan errores, o se hacen cambios en las matrices primarias. Por ello, automatizar la secuencia de análisis hace que frente a estas situaciones todo se puede repetir de un solo *click*.

6 - Calidad y flexibilidad de las salidas gráficas

Uno de los aspectos más llamativos y explotados de R es que provee de herramientas interactivas de análisis con una significativa flexibilidad en la visualización de datos y la producción de gráficos de alta calidad para publicaciones académicas (Mair *et al.,* 2015). Existe una amplia serie de paquetes como *ggplot2* y *plotly* que permiten el desarrollo de gráficos de elevada estética y profesionalidad.

Este es uno de los aspectos más alabados y que ubican a R por encima de cualquier otro sistema de programación de Ciencia de Datos. De hecho, existe toda una nueva generación de representaciones gráficas asociadas al desarrollo de R, que anteriormente estaban limitadas a unos pocos investigadores con facilidades para el diseño gráfico y el dominio de otros programas (*CorelDraw, Photoshop*). Todas las funciones para las representaciones pueden ser modificadas para ajustarlas a las necesidades o gustos específicos de cada autor.

Esta ventaja es ejemplificada en la figura 3, donde se muestra un gráfico de dispersión típico hecho en Excel y se compara con las variantes realzadas que se generan con la aplicación de R *Extended Scatter Graphics* que puede ser ejecutada *online* (https://vmra.shinyapps.io/bivariados/) o descargada gratuitamente para su ejecución local. Esta aplicación, con una interfase web mucho más amigable que el entorno de programación de R fue desarrollada en el propio lenguaje R, lo cual también es una ventaja para los interesados en crear herramientas para ayudar a aquellos investigadores menos hábiles en la codificación.

7 - R permite el desarrollo de aplicaciones para el entorno web

Como lenguaje de programación, R permite el desarrollo de aplicaciones web a través del paquete *Shiny*. El código mencionado anteriormente para facilitar la creación de gráficos de dispersión es un ejemplo de ello (Fig. 4) y puede ser utilizado sin conocimiento previo de programación en R. Con esta herramienta se pueden implementar métodos y herramientas de acceso online o de ejecución local y así facilitar el trabajo de personas con menos habilidades para la codificación o crear recursos educativos para la enseñanza de métodos de trabajo. Por ejemplo, en la dirección URL https://saskiaotto.de/shiny/ pueden explorarse varias aplicaciones desarrolladas en la *Duke University* para estudiar aspectos estadísticos.

Otros ejemplos de aplicaciones con interfaces web para investigaciones, generadas en R, son el paquete Wallace para modelos de nicho ecológico (Kass *et al*., 2018), el paquete rapidPop para monitorear vida silvestre con cámaras trampa (Tabak *et al*., 2020), la aplicación DAME (*Dynamic Assessment of Microbial Ecology*) (Piccolo *et al*., 2018), entre otras. En el contexto nacional, los autores de este trabajo han experimentado con el desarrollo de *FoliométriK* (Ramírez-Arrieta y Denis, 2020), una aplicación para automatizar la toma de mediciones morfométricas de hojas a partir de fotografías calibradas y su análoga *OvometriK* para morfometría de huevos de aves (Ramírez-Arrieta *et al.,* 2020). Además, han desarrollado dos aplicaciones para la obtención de figuras estadísticas con las facilidades de R pero sin necesidad de conocer el lenguaje de programación: *Extended Scatter Graphics*, presentada en este artículo y *Extended Boxplot Graphics* (Denis y Ramírez-Arriaga, 2020).

8 - Existe una amplia base de literatura, libros y recursos para el aprendizaje de R

Otra de las ventajas del lenguaje R es que existen numerosos libros, artículos y recursos en línea disponibles que permiten su aprendizaje autónomo. De hecho, muchos de los libros actuales de bioestadística desarrollan sus ejemplos en R y proveen los códigos y datos para desarrollar los ejercicios (ej.: Smith y Warren, 2019; Faraway, 2016; Zuur *et al.,* 2013; Palacio *et al.,* 2020).

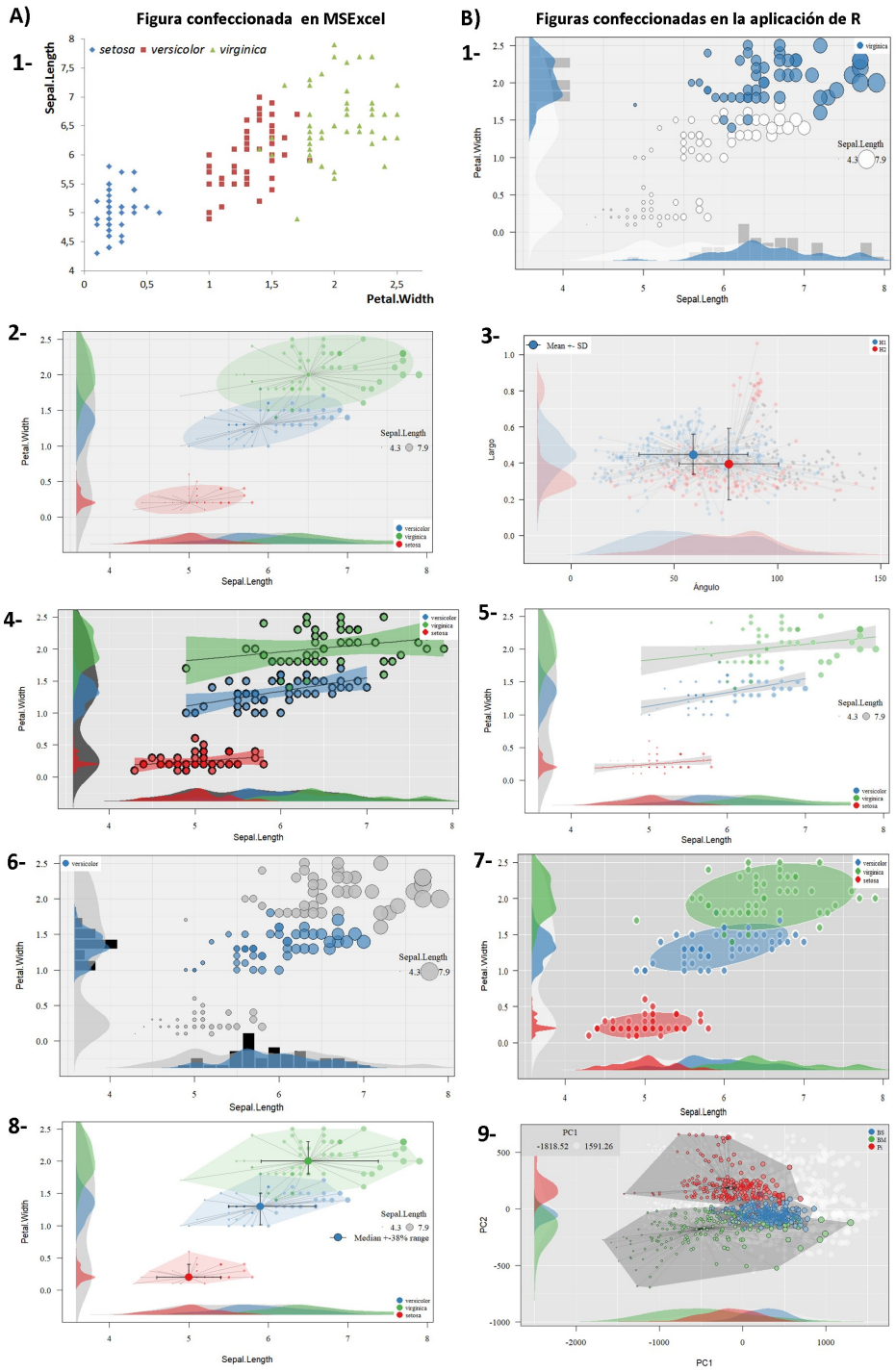

Figura 3: Comparación de un gráfico de dispersión a partir del conjunto de datos Iris creado en Excel (A) y comparación con las posibilidades de mejoras que ofrece R (B) implementadas en la aplicación *Extended Scatter Graphics*, desarrollado por el coautor de esta comunicación y disponible para su uso online o descarga en: https://vmra.shinyapps.io/bivariados/.

Figure 3: *Comparison of a scatterplot created in Excel using Iris dataset and its enhanced possibilities using the R script Extended Scatter Graphics written by the coauthor of current communication and freely available for online use or to download and local use from https://vmra.shinyapps.io/bivariados/.*

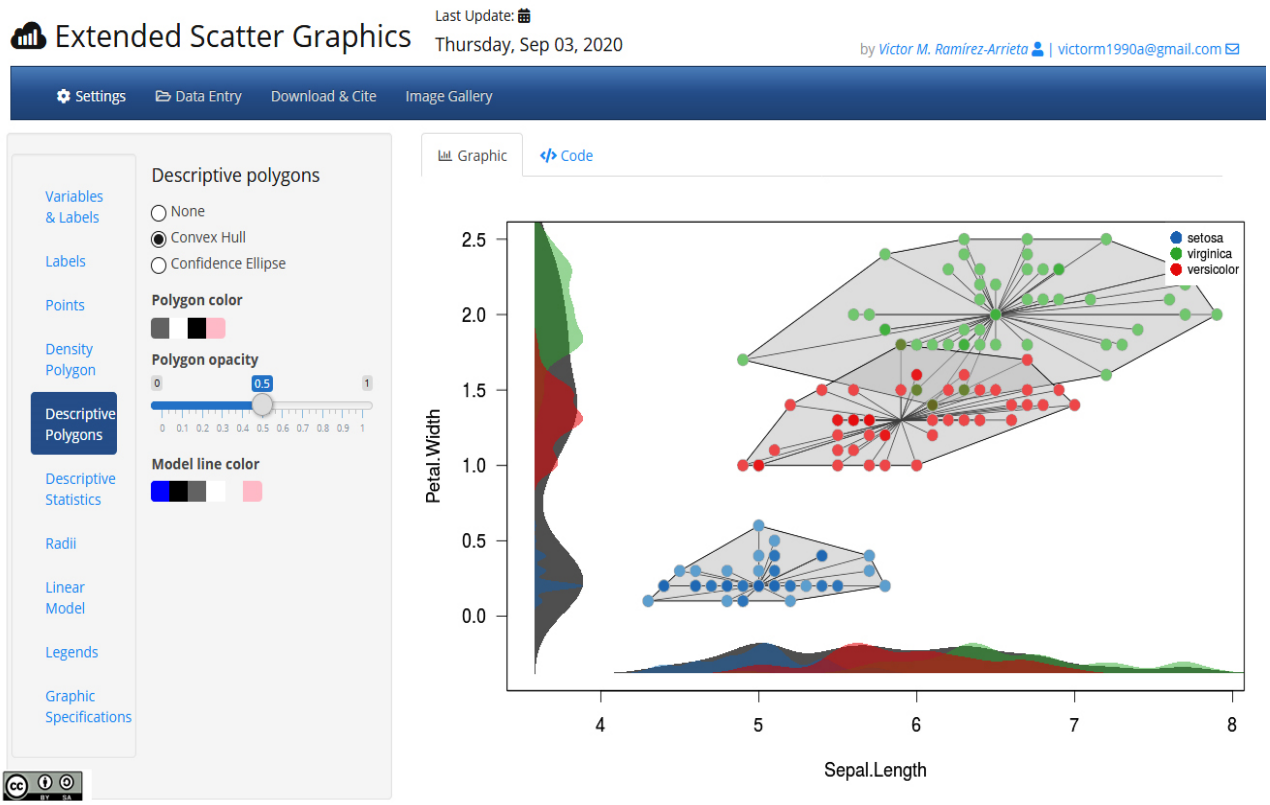

Figura 4: Pantalla de la interfase web que utiliza el script de R de la aplicación *Extended Scatter Graphics*, utilizable online en https://vmra.shinyapps.io/bivariados/ o descargable para uso local y que permite la generación de figuras de dispersión con las opciones de mejoras que posibilita el R.

Figure 4: *Screen capture of the web interface used in R script of the app Extended Scatter Graphics, available online in https://vmra.shinyapps.io/bivariados/ or downloadable for local use, for creating advanced scatter plot with options for enhancement possible with R.*

9 - R tiene una activa comunidad de usuarios en línea

La comunidad de usuarios de R en línea es enorme y con ellos es posible aclarar dudas con rapidez. Existen múltiples grupos y sistemas de comunicación que facilitan este intercambio de ideas, la resolución de problemas y el intercambio de código para especificaciones y desafíos, por ejemplo: *R-sig*, *StackOverflow*, *DataFlair*, *R-Blogs*. Esta comunidad continuamente se mantiene empleando y chequeando los nuevos códigos, con la costumbre general de reportar y corregir rápidamente los errores que se puedan detectar, en una forma avanzada de curación colectiva de la información. Se ha planteado también que R promueve la comunicación entre ecólogos y estadísticos (Andy, 2015; Higgs, 2015).

10 - Facilita el manejo de cantidades masivas de datos: soluciones Big-Data

R es particularmente útil para investigaciones que generan grandes cantidades de datos. Con frecuencia los estudios experimentales solo generan pocos resultados cuantitativos, que son analizados con métodos tradicionales de estadística frecuentista básica y en estos casos, no hay mucha relevancia en el software que se emplee. Pero en la actualidad, la bioinformática, la ecoinformática y las ciencias ambientales se enfrentan cada vez más a cantidades masivas de información. Como programa informático no requiere de un poder computacional muy alto, siendo fácilmente manejables matrices muy grandes, con las cuales otros programas como Excel se demoran excesivamente o se

bloquean con facilidad. Además, con R se hace posible la computación distribuida en línea (*on-cloud computing*). Esta se refiere al empleo de servicios de la nube para desarrollar tareas complejas o que requieren mucha información, de manera que se desarrolla un procesamiento de manera simultánea y dividida entre múltiples nodos de procesamiento, con lo cual se reduce el tiempo de ejecución al analizar grandes bases de datos.

De hecho, en el entorno virtual para investigaciones *D4Science*, desarrollado por la Comunidad Económica Europea, existe una versión online de R totalmente funcional (https://i-marine.d4science.org/web/rstudiolab), que emplea los recursos de redes de computo *online* sin depender de una instalación local o de los recursos de memoria y velocidad, generalmente limitados, de las computadoras personales.

11 - R tiene un papel importante en la batalla por la replicabilidad de las investigaciones publicadas

En la actualidad se ha reconocido la existencia de una crisis global en la reproducibilidad y replicabilidad de las publicaciones científicas (Denis, 2020). En Ecología, la replicabilidad muchas veces está limitada por las diferencias de contextos y las variaciones temporales de los procesos, por lo que se requiere un esfuerzo extra para asegurar al menos la reproducibilidad. Para ello, no solo es importante el acceso abierto a los datos primarios empleados para los análisis sino también vencer los desafíos que conllevan los complejos métodos utilizados para manejar diversos datos ecológicos (Hampton *et al.,* 2015; Roche *et al.,* 2015; Lortie 2017; Lowndes *et al.,* 2017). Para esto R es una herramienta ideal, ya que puede emplearse para describir y automatizar el flujo de trabajo y repetir las decisiones en el tratamiento de las matrices de datos.

Al trabajar con grandes matrices de datos, se desarrolla un proceso de limpieza inicial que incluye: ordenamientos, filtrados, cálculo de variables derivadas, trabajo con datos faltantes o valores atípicos, mezcla de matrices, transposiciones… Este proceso puede tener una influencia fuerte en el resultado final, pero raramente es descrito en suficiente detalle como para ser reproducido por otro investigador.

12 - Codificar en R y poner los datos disponibles puede facilitar la publicación en revistas importantes

Cada vez más, las revistas de ecología están solicitando que los autores hagan públicos sus códigos conjuntamente con los datos (Barnes, 2010; Nosek *et al.,* 2015; Mislan *et al.,* 2016). Esto mejorará la capacidad de revisar los trabajos y comunicar con precisión todos los componentes asociados con el flujo de la investigación. Este énfasis de las revistas está aparejado con el desarrollo de repositorios que permiten compartir y archivar códigos (ej.: Dríada, GitHub y Zenodo), con lo cual se está gestando un cambio en la cultura de colaboración científica que promueve un avance más rápido y seguro en esta ciencia (Hampton *et al.,* 2015).

Hay un grupo de revistas del primer cuartil, relacionadas con el campo de la Ecología y enfocadas en nuevas herramientas y aplicaciones, en las cuales más de la mitad de las publicaciones ya utilizan y comparten códigos de R: *Methods in Ecology and Evolution*, *Global Ecology and Biogeography*, *Ecography*. Les siguen otras revistas en las cuales más de la tercera parte de los artículos emplean este programa: *Diversity and Distributions*, *Journal of Ecology*, *Journal of Animal Ecology*, *Oikos*, *Ecology*, *Journal of Applied Ecology* y *Ecology Letters* (Lai *et al.,* 2019).

Nada es perfecto… ¿Cuáles son las desventajas?

Por supuesto, como cualquier herramienta, no todo es positivo con R. La curva de aprendizaje es mucho más lenta, como siempre sucede con los lenguajes de programación que conllevan escribir código, frente a la intuitiva interacción con entornos visuales en los cuales los botones, íconos y menús facilitan un rápido aprendizaje para un uso básico. Muchos investigadores, sobre todo los que tienen menos dominio del entorno digital, prefieren los programas de interface interactiva (de tipo *click-and-go*) porque no requieren de aprender a programar. El pensamiento estructurado de un programador no lo tiene todo el mundo de forma instintiva, y aun así es muy importante para un científico. Para llegar a niveles profundos de dominio de R se necesita invertir tiempo y esfuerzo de aprendizaje, aunque las operaciones básicas más sencillas son rápidas de aprender.

Incluso la tarea de encontrar los paquetes más apropiados para una tarea puede consumir bastante tiempo. Al existir tantos paquetes la búsqueda de funciones específicas puede ser compleja. Para ello existen numerosas herramientas para ello: el propio CRAN tiene un directorio que ordena los paquetes por áreas o temáticas (*package Task Views*, https://cran.r-project.org/web/views/) y el buscador de *Microsoft R Application Network* (https://mran.microsoft.com/packages/) facilita las búsquedas. El sitio *Rdocumentation.org* provee la lista de paquetes ordenados por su popularidad (dada por el número de descargas (https://www.rdocumentation.org/trends) y los paquetes más utilizados en *GitHub* pueden encontrarse en el *Trending R repositories list* (https://github.com/trending/r?since=monthly). Existe un paquete llamado *packagefinder* que, desde el propio entorno de R accede a CRAN y busca otros paquetes según palabras clave o autores (Zuckarelli, 2019). Pero esta multiplicidad de herramientas hace que el aprendizaje de R conlleve aprender y familiarizarse con muchas otras herramientas modernas como el propio *GitHub*.

Otra de las desventajas reconocidas de R es que, durante el manejo de grandes matrices de datos en computadoras personales puede ser más lento que otros programas, ya que requiere que los datos estén almacenados en la memoria física. O sea, el programa no controla el manejo de la memoria RAM por lo que al ejecutarse códigos demandantes puede llegar a consumir toda la memoria disponible, enlenteciendo otras aplicaciones simultáneas. Esto sucede, sobre todo, con algoritmos pobremente programados y aunque hay maneras de optimizarlo por medio de paquetes como *pqR*, *renjin*, *FastR*, *Riposte* y otros, continúa siendo un reto. Esta desventaja puede ser sobrellevada si se emplean recursos de la nube (*on-cloud computing*).

Otra desventaja es el riesgo del uso de paquetes sin suficiente prueba o documentación detallada, aunque los repositorios oficiales la requieren para poderse publicar el paquete y existe la posibilidad de revisar el código antes de usarlo. Una última desventaja es que, al ser gratuito, no existe un sistema centralizado de atención al consumidor con el cual quejarse si algo no funciona bien, aunque la extensa comunidad *online* funciona muy bien en este rol.

¿Cómo podría potenciarse el uso de R en las investigaciones cubanas?

La práctica profesional es una extensión de lo que se aprende durante la formación de pregrado. De hecho, no es descabellado asumir que la generalización del uso del programa estadístico *Statistica* en Cuba se debe a su introducción en la docencia de la carrera Biología desde el plan de estudios C perfeccionado y su promoción y utilización en actividades docentes de varias asignaturas de pregrado. Por ello, la introducción de R en el currículo universitario debe ser el primer paso. Dada la universalidad de su aplicación con independencia de la rama de la Biología (desde las ramas moleculares a la ecología), no debería ser tan solo una asignatura optativa sino incorporarse como asignatura básica.

En términos de eficiencia, en medio de una tendencia a la reducción de la carga horaria tal vez debería replantearse la asignatura de computación de la disciplina de Matemática y computación, y enfocarse en este lenguaje de programación. Dentro del sistema de objetivos de esta asignatura está el de diseñar algoritmos de problemas sencillos y realizarlos a través de un lenguaje de programación y entre las habilidades que debe formar están la capacidad de utilizar software para automatizar el procesamiento de los datos y realizar algoritmos sencillos por medio de un lenguaje de programación, comprobando la solución del problema con datos conocidos. Esta asignatura se imparte en el tercer semestre, conjuntamente con las asignaturas de Ecología y Bioestadística. La inclusión de R como programa de estudio puede abrir posibilidades muy grandes de trabajo interdisciplinar, así como el desarrollo de actividades prácticas y evaluaciones conjuntas, con actividades que se complementen entre asignaturas. En este sentido, un cambio de perspectiva se impondría, desde la anterior de considerar a la matemática y la estadística como bases para la programación hacia una en la cual la programación se convierte en herramienta práctica para el aprendizaje significativo de conocimientos matemáticos y estadísticos.

De manera inmediata, el desarrollo de programas de postgrado y la presión colectiva desde las revistas científicas para potenciar el uso de esta herramienta también pueden ser un buen aporte. De hecho, desde el año 2015, el Instituto Nacional de Salud (NIH) de los EEUU ha promovido la inclusión de entrenamientos para desarrollar habilidades de programación en los programas de postgrado de disciplinas biomédicas.

¿Por qué R y no *Phyton*?

Una pregunta válida sería el por qué no aprender *Phyton* en lugar de R. Conjuntamente con R, *Phyton* es el otro de los programas más utilizados en el campo de la Bioinformática, tanto como herramienta para su uso en investigaciones o para desarrollar nuevas aplicaciones. La principal diferencia entre ambos es que *Phyton* es un lenguaje para propósitos generales mientras que R fue desarrollado por estadísticos y por ello maneja el lenguaje técnico de esta especialidad.

Phyton tendría como única desventaja el que no tiene la gran cantidad de paquetes como R, aunque lo supera para aplicaciones de computación distribuida, cuando debe interactuarse con servidores remotos o *clusters* de computadoras o para crear aplicaciones compiladas para uso independiente. Otros lenguajes de programación como *Pearl*, C y C++ son también válidos, pero son mucho menos usados en el campo profesional de los biólogos, son más difíciles de aprender, más enrevesados y necesitan mucho más código para ejecutar las mismas funciones que en R o *Phyton*. Para aplicaciones basadas en web son útiles *Ruby*, *Java*, *JavaScript* o *PHP*, pero tienen menos aplicaciones para investigaciones biológicas.

Si el objetivo de las asignaturas de programación en las carreras de Biología no es formar programadores generales sino desarrollar habilidades cognitivas relacionadas con el pensamiento estructurado y algorítmico, mientras se dotan a los estudiantes de herramientas útiles para su desarrollo profesional, el lenguaje R tiene las máximas potencialidades. Aunque algunos han llegado a plantear que R no es un verdadero "lenguaje de programación", la mayoría ya acepta que sí lo es ya que con él se pueden enfrentar las mismas tareas que con *Phyton* u otros lenguajes.

A modo de conclusión…

Las aplicaciones en el lenguaje R no muestran ningún signo de disminución a corto o mediano plazo. Es muy probable que en Ecología y Evolución se continúe usando de manera intensa para la investigación, por la creciente disponibilidad de paquetes relevantes para el análisis de datos y las facilidades para que los nuevos métodos se reproduzcan y apliquen. Esto está contribuyendo a un acelerado desarrollo de las investigaciones modernas en Ecología y muchas otras ramas de las ciencias de la vida, y nos conecta directamente a la Ciencia de Datos y al pensamiento computacional (Visser *et al.* 2015). Las investigaciones ecológicas en Cuba deberían hacer un esfuerzo para incorporarse a esta corriente de avanzada.

Es cierto que puede aducirse que las investigaciones más básicas y simples en Ecología son perfectamente desarrollables con los programas tradicionales sin una necesidad imperiosa de aprender a programar en R. Pero el conflicto potencial con las leyes de propiedad intelectual puede en algún momento convertirse en una barrera al empleo irrestricto de copias ilegales de los programas comerciales de análisis más empleados en el momento actual. Cambiar a programas abiertos, de potencias similares o superiores, es una perspectiva estratégica a fomentar. Además, si se tienen en cuenta los profundos cambios que se están dando en el escenario científico por los niveles de informatización, los fenómenos de *Big-data* y *Big-literature*, el desarrollo de la Ciencia de Datos, el uso creciente de métodos de aprendizaje de máquina e inteligencia artificial, la crisis de la replicabilidad y la crisis estadística en las publicaciones, dedicar un esfuerzo al aprendizaje de R puede ser considerada una muy buena inversión para los jóvenes profesionales de la Biología que desee avanzar en la Ciencia.

LITERATURA CITADA

Andy, H (2015). The new statistics with R: an introduction for biologists. Oxford University Press, Oxford, UK.

Barnes, N (2010). Publish your computer code: It is good enough. Nature 467:753.

Bivand, R. S., E. Pebesma y V. Gomez-Rubio (2013). Applied spatial data analysis with R. Second edition. Springer, New York, New York, USA.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, *et al.* (2009). Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24:127–135.

Bollmann, S., D. Cook, J. Dumas, J. Fox, *et al.* (2017). A first survey on the diversity of the R community. R Journal 9:541–552.

Borcard, D., F. Gillet y P. Legendre (2018). Numerical ecology with R. Second edition. Springer, New York, New York, USA.

Cadotte, M. W., T. J. Davies y P. R. Peres-Neto (2017). Why phylogenies do not always predict ecological differences. Ecological Monographs 87:535– 551.

Clarke K.R. y R.M. Warwick (2001). Change in Marine Communities. 2da Ed. PRIMER-E Ltd, Plymoutli.

Denis, D (2020). Las crisis de la ciencia moderna. Rev. Cub. Cienc. Biol. 4(2): 1-16

Denis, D. y V.M. Ramírez-Arriaga (2020). Si una imagen vale 1000 palabras: ¿cuándo puede decir un gráfico de cajas? Rev. J. Bot. Nac. 41: 57-69

Diniz-Filho, J. A. F., T. N. Soares, J. S. Lima, R. Dobrovolski, V. L. Landeiro, *et al.* (2013). Mantel test in population genetics. Gen. Mol. Biol. 36:475–485.

Dixon, P (2003). VEGAN, a package of R functions for community ecology. J. Veg. Sci.14:927–930.

Excoffier, L., P. E. Smouse y J. M. Quattro (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes – application to human mitochondrial-DNA restriction data. Genetics 131:479–491.

Faraway, J. J. (2016). Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. Chapman and Hall/CRC.

Fontenla, J.L., Y. Fontenla, Z. Cuervo y A. Álvarez de Zayas (2019). Red de interacción ecológica insectos-plantas en Playas del Este, la Habana, Cuba. Acta Botánica Cubana 218(2): 129-142

Hammer, Ø., Harper, D.A.T. y P. D. Ryan (2001). PAST: Paleontological Statistics Software Package for Education and Data Analysis. Palaeontologia Electronica 4(1): 9 pp.

Hampton, S. E., SS Anderson, SC Bagby, C Gries ,*et al.* (2015). The Tao of open science for ecology. Ecosphere 6:120.

Higgs, D M (2015). Ecology and statistics: A healthy union? BioScience 65:1021–1025.

Kass, J.M., B. Vilela, M.E. Aiello-Lammens, R. Muscarella, *et al.* (2017). Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. Methods Ecol Evol. 2018, 1–6. https://doi.org/10.1111/2041-210X.12945

Lai, J.; C.J. Lortie, R.A. Muenchen, J. Yang y K. Ma. (2019). Evaluating the popularity of R in ecology. Ecosphere 10(1):e02567. 10.1002/ecs2.2567

Lortie, C. J. (2017. Open sesame: R for data science is open science. Ideas in Ecology and Evolution 10:1–5.

Lowndes, J. S. S., B. D. Best, C. Scarborough, J. C. Afflerbach, M. R. Frazier *et al.* 2017. Our path to better science in less time using open data science tools. Nature Ecology & Evolution 1:160.7

Mair, P., E. Hofmann, K. Gruber, R. Hatzinger *et al.* (2015). Motivation, values, and work design as drivers of participation in the R open source project for statistical computing. Proc. Nat. Acad. Sci. USA 112:14788–14792.

Mislan, K. A. S., J. M. Heer y E. P. White. (2016). Elevating the status of code in ecology. Trends Ecol. Evol. 31:4–7.

Muenchen, R. A. (2017). The popularity of data science software. Disponible en: http://r4stats.com/articles/popularity. Último acceso: 17 de abril de 2020.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, *et al.* (2015). Promoting an open research culture. Science 348:1422–1425.

Palacio, F.X.; M. J. Apodaca y J. V. Crisci. (2020). Análisis multivariado para datos biológicos: teoría y su aplicación utilizando el lenguaje R. 1a ed. Ciudad Autónoma de Buenos Aires. Fundación de Historia Natural Félix de Azara.

Phipson, B., y G. K. Smyth. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly Drawn. Statistical Stat. Appl. Gen. Mol. Biol. 9. https://doi.org/10.2202/1544-6115.1585

Piccolo, B.D., U. D Wankhade, Sree V Chintapalli, Sudeepa Bhattacharyya *et al.* (2018). Dynamic assessment of microbial ecology (DAME): a web app for interactive analysis and visualization of microbial sequencing data. Bioinformatics, 34(6): 1050–1052, https://doi.org/10.1093/bioinformatics/btx686

R Development Core Team. (2017). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramírez-Arrieta, V. M., D. Denis y Y. Ferrer-Sánchez. (2020). Evaluación de un protocolo automatizado para la extracción de medidas morfométricas de huevos de aves a partir de fotografías digitales. Researchgate Preprint. DOI: 10.13140/RG.2.2.11933.33760

Ramírez-Arrieta, V. y D. Denis (2020). Implementación de un paquete de r para mediciones morfométricas automatizadas de hojas a partir de fotografías digitales. Rev. J. Bot. Nac. 41: 15-23

Roche, D. G., L. E. B. Kruuk, R. Lanfear, y S. A. Binning. (2015). Public data archiving in ecology and evolution: How well are we doing? Plos Biology 13:e1002295.

Simpson, G. L. (2018). Modelling palaeoecological time series using generalized additive models. bioRxiv. https://doi.org/10.1101/322248

Smith, C. y M. Warren. (2018). GLMs in R for Ecology. Libro electrónico. https://www.amazon.es/GLMs-Ecology-English-Carl-Smith-ebook/dp/B07WSD46MM/

Swenson, N. G. (2014). Functional and phylogenetic ecology in R. Springer, New York, New York, USA.

Tabak, M. A. J. S. Lewis, P. E Schlichting, N. P. Snow *et al.* (2020). rapidPop: Rapid population assessments of wildlife using camera trap data in R Shiny Applications. bioRxiv 2020.03.30.017103; doi: https://doi.org/10.1101/2020.03.30.017103

Tippmann, S. (2015). Programming tools: adventures with R. Nature 517:109–110.

Touchon, J. C. y M. W. McCoy. (2016). The mismatch between current statistical practice and doctoral training in ecology. Ecosphere 7:e01394.

Visser, M. D., S. M. McMahon, C. Merow, P. M. Dixon, S. Record y E. Jongejans. (2015). Speeding up ecological and evolutionary computations in R; essentials of high performance computing for biologists. Plos Computational Biology 11:e1004140.

Viswam, N. y D K. Srinivasa. (2019). R Programming in Different Fields. IJCSET 9: 1-8

Zuckarelli, J. (2019). packagefinder: Comfortable search for R packages on CRAN directly from the R console. CRAN. Último acceso: 29 de septiembre de 2020.

Zuur, A.F., Hilbe, J.M. y Leno, E.N. (2013). A beginner's guide to GLM and GLMM with R: A frequentist and Bayesian perspective for ecologist

● ● ●