UNIVERSIDAD DE COSTA RICA
ESCUELA DE BIOLOGÍA

PROBLEMAS ESPECIALES EN ECOLOGÍA
Programación y métodos estadísticos avanzados en R

CICLO I- 2017
CREDITOS: 4
ESTUDIANTES: pregrado y posgrado
HORARIO: 4 horas por semana (lunes 9-10:50; jueves 10-11:50)
AULA: L: 220 / J: 201
REQUISITOS: Genética, Estadística I / II y Ecología
PROFESORES:

- Mario Espinoza Mendiola; correo: marioespinozamen@gmail.com; mario.espinoza_m@ucr.ac.cr;
  Página web: www.mespinozamen.com
  Oficina: Biología 203; CIMAR 230;
  Tel: 2511-2208 / 2511-8683
- Marcelo Araya Salas; correo: araya-salas@cornell.edu
  Oficina: Laboratorio de Bioacústica (Aula 170, Biología)

HORAS DE ATENCIÓN: L: 11-12:00 (Biología) / J: 13-15:00 (CIMAR)

DESCRIPCIÓN DEL CURSO

Las ciencias biológicas involucran la colección sistemática, organización, análisis y presentación de la información. En biología, la mayoría de las investigaciones generan una gran cantidad de datos cuantitativos que deben ser analizados de forma apropiada para obtener resultados confiables (Zar 2010, Zuur et al. 2010). En este sentido, la estadística es una herramienta que permite ordenar, analizar e interpretar datos biológicos. Para dicho fin, se han desarrollado una gran variedad de programas estadísticos computacionales que permiten procesar extensas bases de datos de distintos tipos a gran velocidad. Sin embargo, muchas veces estos programas o paquetes estadísticos tienen la limitante de que, son difíciles de adquirir por su alto costo, deben ser actualizados periódicamente, no proveen herramientas para la manipulación de datos, o están restringidos por un número definido de pruebas estadísticas o funciones matemáticas.

Además, debido a que existe un gran número de paquetes estadísticos, a veces es necesario aprender a manipular diferentes programas computacionales (i.e., softwares) para realizar análisis específicos. Afortunadamente, en la última década surgió el sistema computacional R, el cual ofrece una plataforma de acceso libre que le permite al usuario programar sus propias funciones, realizar pruebas estadísticas, graficar, manipular bases de datos extensas y de naturaleza distinta, así como compartir nuevas herramientas desarrolladas independientemente por los usuarios (Crawley 2007, Fuchs & Barrantes 2015).

El desarrollo de la plataforma R para estadística computacional esta grandemente influenciado por la idea de “fuentes de libre acceso”: La distribución base de R y un gran número de contribuciones (extensiones) están disponibles bajo los términos de Fundación de Software Libre (R Development Core Team 2014). Esta licencia tiene dos implicaciones importantes para analistas de datos que trabajan con R. Primero, todos los códigos completos están disponibles y son de acceso libre para todos, por lo que

cualquier persona puede examinar, editar y manipular el cogido de acuerdo a sus necesidades. Segundo, esta plataforma está constantemente creciendo, renovándose y actualizándose de acuerdo a las necesidades específicas de los usuarios. Esto ha hecho que R se haya convertido en una de las plataformas más populares y utilizadas en una amplia variedad de disciplinas (Crawley 2007, R Development Core Team 2014).

Por esta razón, el manejo del ambiente de programación R resulta una herramienta fundamental para el desarrollo profesional de los estudiantes de ciencias biológicas, independientemente de su área de estudio. La carrera de biología ofrece cursos introductorios de estadística donde los estudiantes utilizan R para realizar diversas pruebas estadísticas. Además, algunos cursos introducen a sus estudiantes en aplicaciones específicas de R en sus disciplinas (ej. paquetes de análisis comparativos en el curso Análisis Comparativo Filogenético, paquetes de análisis espaciales y multivariados en el curso de ecología y conservación de tiburones y rayas). Sin embargo, muchos estudiantes de la carrera de Biología: (1) siguen teniendo una dependencia muy grande hacia programas básicos como Excel para manipular bases de datos; (2) continúan usando softwares estadísticos comerciales de interfaz gráfica de usuario (no programables) para sus análisis; (3) o circunscriben sus análisis estadísticos a las herramientas disponibles en softwares comerciales, limitando de manera importante las opciones de análisis.

Este curso pretende profundizar en los elementos de programación computacional, manipulación de bases de datos, diseño experimental, graficación personalizada, y el uso de técnicas y modelaje estadístico avanzadas utilizando R como plataforma (Touchon & McCoy 2016, Warton et al. 2016). El curso está dirigido a estudiantes avanzados de carrera, de licenciatura o postgrado. En el curso se pretende cubrir las bases de los principales análisis y técnicas de modelaje estadístico, así como análisis emergentes. Todas las semanas se realizarán prácticas de laboratorio en donde se aplicarán los conceptos desarrollados durante las clases de teoría. Además, durante el curso cada estudiante presentará un paquete o extensión (conjunto de herramientas aplicables a análisis específicos) de R, en donde profundizará sobre sus aplicaciones en el campo de la biología demostrando en clase en que consiste el análisis.

OBJETIVO GENERAL

- Enseñar a estudiantes elementos básicos y avanzados de programación computacional que les permitan manipular, analizar, graficar e interpretar información biológica utilizando R como plataforma de trabajo.

OBJETIVOS ESPECIFICOS

- Familiarizar al estudiante con la programación en R.
- Brindar herramientas para la manipulación de bases de datos usando la plataforma R.
- Emplear métodos de visualización de datos usando la plataforma R.
- Cubrir algunos de los análisis estadísticos tradicionales usando la plataforma R.
- Realizar análisis de modelaje usando la plataforma R.

- Proveer a los estudiantes con experiencia en la aplicación de las herramientas brindadas por medio de prácticas y proyectos de investigación.

CONTENIDO Y CRONOGRAMA

- Periodo del 13 de marzo al 7 de abril, 2017
  - Introducción al curso: Discusión del programa y descripción de la evaluación del curso.
  - Introducción a la programación en R: Familiarizar al estudiante con la programación en R. En este tema se abordarán las bases de programación en R:
    - ✓ Uso de funciones básicas y avanzadas para crear, importar y manipular bases de datos (paquetes "dplyr", "data.table", y "reshape").
    - ✓ Combinar bases de datos, cambiar formatos de bases de datos (largo vs. ancho)
    - ✓ Métodos de visualización en R
  - Reportes dinámicos, elegantes y flexibles en R con las extensiones "knitr" y "Rmarkdown". Durante este periodo se generarán reportes dinámicos, elegantes y flexibles en R que permitan al estudiante distribuir, comunicar y divulgar su ciencia.

- **Semana santa (10-14 de abril): No hay lecciones**

- Periodo del 17 al 21 de abril, 2017
  - Programación avanzada en R: Fortalecer las bases de programación básica en R con elementos más complejos. Para esto se abordarán temas que incluyan el uso de operadores y funciones lógicas (expresiones "if", "ifelse", "for", "while", "Xapply", etc. .) y de métodos para mejorar el desempeño de rutinas computacionalmente intensivas (e.g. paralelización) Durante este periodo, también se le enseñará al estudiante a interpretar y escribir funciones personalizadas que pueden ser de gran utilidad en sus disciplinas, dentro de la biología.

- **Semana Universitaria (24 al 28 de abril, 2017): No hay lecciones**

- Periodo del 1 al 12 de mayo, 2017
  - Graficación avanzada en R: Emplear métodos de visualización de datos usando la plataforma R. En esta sección se cubrirán los principales métodos de graficación en R, incluyendo el uso de la librería "GGPLOTS2", el cual permite generar gráficos más atractivos y personalizados para publicaciones científicas.

- Periodo del 15 de mayo al 16 de junio, 2017
  - Análisis estadísticos en R: Durante este periodo se cubrirán las bases de algunos de los análisis estadísticos tradicionales y análisis más complejos usando la plataforma R. Esta sección

pretende darle las herramientas a estudiantes de biología para que puedan analizar algunas de las pruebas estadísticas más utilizadas en R.

- Entre los análisis que se van a cubrir en el curso están: Análisis de Frecuencias (Chi-cuadrado, Pruebas de Bondad de Ajuste, Tablas de Contingencia, "odds ratio", etc.), t-student, Análisis de Varianza ("ANOVA"), Análisis de Regresión Lineal y Correlación, Análisis de Regresión Múltiple, Modelos Lineales Generalizados ("GLMs", por sus siglas en inglés), Modelos Lineales Generalizados Mixtos ("GLMMs", por sus siglas en inglés), Modelos Aditivos Generalizados ("GAMs", por sus siglas en inglés), Análisis Multivariados de ordenación ("nMDS – non-metric Multidimensional Scaling", "Cluster Analysis", "CCA – Canonical Correspondence Analysis", "PERMANOVA – Permutational MANOVA"), Aleatorizacion y remuestreo (bootstraping).

- **Entrega de primer examen parcial: jueves 18 de mayo, 2017**

- Periodo del 19 al 30 de junio, 2017
  - Presentaciones y trabajos finales de estudiantes.

- **Entrega de segundo examen parcial: jueves 29 de junio, 2017**

Metodología y actividades para cumplir con los objetivos

- Clases prácticas con computadora y participación activa de estudiantes
- Tareas
- Presentación y discusión de librerías en R
- Exámenes para resolver en la casa
- Proyecto final para el curso

EVALUACIÓN

El curso incluirá cuatro rubros a evaluar:

1- Dos exámenes parciales (20%c/u).

2- Tareas y participación (25%): Durante el semestre se dejarán 4 ejercicios de tarea (5%c/u), que le permitirá al estudiante poner en práctica lo aprendido en clase. Las tareas consistirán en ejercicios de: (i) manipulación de bases de datos, (ii) graficación, (iii) uso de funciones y operadores lógicos, y (iv) análisis estadísticos aplicados, en los cuales el estudiante deberá mostrar sus habilidades para resolver una situación particular. Las tareas deberán ser entregadas una semana después de su asignación, empleando un reporte dinámico generado por medio de los paquetes "knitr" y "Rmarkdown" para asegurar su reproducibilidad.

3- Participación en clase de los estudiantes (5%).

4- Presentación y discusión de un paquete (30%): A lo largo del semestre cada estudiante será

responsable de escoger un paquete de R, sobre el cual realizará una presentación breve (10 min max.) seguida de una demonstración con un set de datos ficticio o real.

5- Proyecto Final (25%): Cada estudiante seleccionará un set de datos, y deberá entregar al final del semestre, una pequeña descripción del diseño experimental que utilizó, los análisis estadísticos empleados, y la representación gráfica e interpretación de los resultados generados en R. Este proyecto deberá ser entregado con su respectivo set de datos y el reporte dinámico generado por medio de los paquetes "knitr" y "Rmarkdown" para asegurar su reproducibilidad.

BIBLIOGRAFÍA

Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 59:390–412

Bates D, Maechler M, Bolker BM, Walker S (2014) lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. http://cran.r-project.org/package=lme4.

Bolker BM (2008) Ecological Models and Data in R. Princeton University Press, London

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24:127–35

Borcard D, Gillet F, Legendre P (2011) Numerical Ecology with R. Springer, London

Braak CJF ter, Verdonschot PFM (1995) Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquat Sci 57:255–289

Crawley MJ (2007) The R Book, 2nd Editio. Wiley, Southern Gate

Fuchs EJ, Barrantes G (2015) El lenguaje estadístico R aplicado a las ciencias biológicas. Editorial de la Universidad de Costa Rica, San José, Costa Rica

James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning with Applications in R. Springer

Legendre P, Oksanen J, Braak CJF ter (2011) Testing the significance of canonical axes in redundancy analysis. Methods Ecol Evol 2:269–277

Maindonald J, Braun WJ (210AD) Data Analysis and Graphics Using R - an Example-Based Approach (J Maindonald and W. Braum, Eds.), Third edit. Cambridge University Press, London

R Development Core Team (2014) R: a language and environment for statistical computing.

Reimann C, Filzmoser P, Garrett RG (2008) Statistical Data Analysis Explained.

Touchon JC, McCoy MW (2016) The mismatch between current statistical practice and doctoral training in ecology. Ecosphere 7:e01394

Warton DI, Lyonsy M, Stoklosa J, Ivesz AR (2016) Three points to consider when choosing a LM or GLM test for count data. Methods Ecol Evol:n/a-n/a

Zar JH (2010) Biostatistical Analysis. Prentice Hall, New Jersey

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol 1:3–14

Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM (2009) Mixed Effects Models and Extensions in Ecology with R. Springer, New York, NY

Zuur AF, Saveliev AA, Ieno EN (2014) A beginner's Guide to Generalised Additive Mixed Models with R. Highland Statistics Ltd., Newburgh, United Kingdom

Recursos adicionales en internet

- https://cran.r-project.org
- https://cran.r-project.org/manuals.html
- https://cran.r-project.org/other-docs.html
- https://www.r-project.org/doc/bib/R-books.html
- http://research.stowers-institute.org/efg/R/Color/Chart
- http://research.stowers-institute.org/efg/R/Color/Chart/ColorChart.pdf
- http://www.stat.ubc.ca/~jenny/STAT545A/block14_colors.html
- http://research.stowers-institute.org/efg/R/Graphics/Basics/mar-oma/index.html
- http://www.r-bloggers.com/setting-graph-margins-in-r-using-the-par-function-and-lots-of-cow-milk
- https://www.stat.auckland.ac.nz/~paul/Talks/Rgraphics.pdf
- http://www.statmethods.net/advgraphs/parameters.html
- https://www.r-bloggers.com/a-fast-intro-to-plyr-for-r/
- https://www.r-statistics.com/tag/visualization/
- http://blog.revolutionanalytics.com/2014/04/some-r-resources-for-glms.html
- https://www.datacamp.com/community/tutorials/r-tutorial-apply-family