Universidad de Caldas
Plan Institucional de Actividad Académica
Asignatura de Postgrado

1. Ficha Técnica de la Asignatura

Departamento que oferta		**Ciencias Biológicas**	
Nombre actividad académica		Modelos multinivel en R con énfasis en ecología y evolución	
Código actividad académica			
Palabras clave		Evolución, filogenia	
Prerrequisitos (haber cursado y aprobado)		Bioestadística	
Co-requisito (al menos estar cursando)			
Tipo de asignatura		Electiva (teórico-práctica)	
Fecha de aprobación (AAMMDD)			
Número de acta Consejo Facultad			
Número de créditos que otorga	4	Horas prácticas	24
Horas teóricas	24	Porcentaje de práctica	50
Horas totales presenciales con docente	48	Nota aprobatoria relacion 1:3 g7h0247	3,5
Horas no presenciales	144	Horas de inasistencia para reprobar (sin excusa)	8
Relación presencialidad/no-presencial	1:2	Horas de inasistencia para reprobar (con excusa)	12
Habilitable	No	Cupo máximo de estudiantes	10
Homologable	No	Duración en semestres	1
Validable	No	Duración en semanas	16

JUSTIFICACIÓN

Sobre la necesidad de un análisis de datos con una perspectiva en ecología y evolución

Entender, manejar y analizar datos en ecología y evolución es una de las mayores fortalezas que un investigador puede lograr. Este logro requiere un proceso de mucha dedicación. En nuestro medio (*e.g.*, muchas universidades a nivel de pregrado e incluso posgrado) nos indican el camino a seguir por medio de cursos básicos de estadística (con un enfoque estrictamente numérico-matemático que considera poco las interpretaciones biológicas) que, comúnmente, limitan sus alcances al *mundo normal*, es decir, datos de distribución gaussiana. Con elementos sobre esta familia de distribución se producen actualmente gran parte de los trabajos de grado en pregrados de ecología, biología y áreas afines. Sin embargo, el camino hacia un manejo y análisis de datos biológicos más apropiado está lejos de allí y las pautas que nos provee nuestro medio para encontrar ese camino son muy limitadas. Consecuentemente, jóvenes investigadores pierden la ruta cuando enfrentan datos no normales, y que, por cierto, son comunes en la naturaleza (*e.g.*, distribuciones de tipo *binomial* y *poisson*, las cuales provienen de presencias y ausencias o conteos, entre otras). Esta pérdida de ruta no solo genera stress en los estudiantes, sino que también los lleva a solucionar sus limitaciones bio-estadísticas por medio de análisis inapropiados. Por ejemplo, es común observar que la principal solución frente a datos que no siguen una distribución normal es el uso de *tests* no paramétricos, que, en casi todos los casos, presentan serias limitaciones. Y además como ocurre típicamente, los resultados de estos *tests* (como también ocurre con los ANOVAS o ANCOVAS) se enfocan principalmente a examinar los valores p (*p-values*) pero no las diferencias entre covariables y su relevancia biológica (por ejemplo, el tamaño de los efectos y su magnitud no son tradicionalmente cuantificados). En adición, luego de aplicar estos análisis clásicos, la alternativa principal empleada son los *tests post*-hoc; sin

embargo, este tipo de *tests* tienen la desventaja de incrementar la probabilidad de cometer errores tipo I (*i.e.* reportar un efecto cuando no lo hay).

El camino más apropiado a seguir, y que se ha posicionado poderosamente desde hace algunas décadas, es el uso de los modelos multinivel (también denominados modelos mixtos, jerárquicos o *GLMMs*). Su fortaleza radica en la flexibilidad para analizar datos de distribución no normal y datos que no son independientes entre sí (un problema casi inherente en todos los datos obtenidos en estudios de campo). Dentro de la familia *GLMMs* es fácil realizar tradicionales t-*tests*, anovas, ancovas, regresiones múltiples, y test no paramétricos (*e.g.*, *Kruskall-Wallis, Mann-Whitney*). Específicamente, para aquellos que estudian ecología del comportamiento, los modelos mixtos permiten calcular fácilmente repetibilidad (*e.g.*, variabilidad dentro de grupos y entre grupos), la cual es clave en esta área de la ecología. Por último, es posible incluir las filogenias de las especies objeto de estudio, de ser el caso, en *R* y ser, como en el caso anterior, integradas en los modelos mixtos.

Sobre la importancia de usar el software R

R (ver: r-project.org) es un programa que en las últimas décadas se ha posicionado ampliamente. Una de las principales ventajas de *R* es ser gratuito. Su funcionalidad es versátil ya que los usuarios usan paquetes con diversas funciones; por ejemplo, en la actualidad existen más de 2000 paquetes que se pueden integrar en *R*; así nuevos avances estadísticos pueden ser rápidamente implementados. Es además ventajoso que estos paquetes vienen con manuales/tutoriales gratuitos que se pueden obtener en línea. Una de las principales críticas de *R* es que se basa en un lenguaje de códigos (*scripts*). No obstante, es un lenguaje que puede ser aprendido y perfeccionado con práctica, y en este contexto la plataforma *Rstudio* puede ser de gran utilidad. *Rstudio* es también gratuito y más interactivo que la consola tradicional de *R*. *R* y *Rstudio* se pueden instalar en Windows, MacOS y Linux, así que no existe ninguna barrera para su implementación. Como se mencionó su principal limitación es el lenguaje de códigos, pero es más sencillo de lo que parece. En otras palabras, una vez se adquieren algunas competencias básicas de *R*, el lenguaje deja de ser un factor negativo y se vuelve positivo. En conclusión, la capacidad de *R* es virtualmente ilimitada. Por ejemplo, para manejar bases de datos de gran tamaño de forma eficiente, para producir gráficos/figuras de alta calidad y para poder replicar los análisis tantas veces como sea necesario (lo cual no es posible con otros programas que no se basan en scripts).

Si bien, actualmente existe una gran diversidad de programas estadísticos, en muchos casos, estos programas requieren del pago de licencias para su uso (*e.g.*, *IBM SPSS, SAS*); y adicionalmente, estos programas solo cubren algunos componentes/temas estadísticos (*e.g.*, el ajuste de modelos se hace bajo un método específico). La popularidad de estos programas se basa en que presentan una interface amigable con los usuarios; es decir, estos programas despliegan por medio de sub-secciones las opciones que el usuario busca/requiere. No obstante, esta ventaja aparente tiene una sería limitación toda vez que la capacidad de análisis termina cuando las sub-secciones del programa terminan.

Consecuentemente, el uso de *R* ha venido en ascenso, particularmente dentro del contexto colombiano, pero aún es común una percepción de temor sobre el programa y consecuentemente, en cierta medida, ha sido excluido en muchos contextos (*e.g.*, algunas universidades locales no fomentan su uso). Un usuario novato frente a *R*, generalmente se desmotiva al no poder ingresar apropiadamente sus datos en la plataforma. Sin embargo, una vez los datos se presentan adecuadamente en *R*, es fácil usar y entender, no solo sus funciones básicas y paquetes asociados sino también aprovechar la versatilidad de *R* en otros contextos, de acuerdo con lo mencionado previamente. Es importante enfatizar que en Colombia ha existido poca oferta sobre cursos sobre el manejo de *R*. Particularmente existe una ausencia de cursos enfocados en hacer un recorrido sobre las principales técnicas y análisis de datos en ecología y evolución, teniendo *R* como eje central. Por ejemplo, el análisis e interpretación de datos biológicos vía modelos multinivel en *R* se ha convertido en el *gold standard* en análisis e interpretación de datos biológicos. Consecuentemente, un curso de maestría y doctorado en *R* con énfasis en ecología y evolución es novedoso y pertinente.

OBJETIVO GENERAL

Realizar un curso teórico-práctico en el software *R* que tiene como objetivo general fomentar un apropiado y eficiente manejo y análisis de datos biológicos con énfasis en modelos multinivel (modelos mixtos).

OBJETIVOS ESPECÍFICOS

- Fortalecer el manejo del software *R*
- Realizar un manejo eficiente de datos por medio de R y de su paquete *dplyr.*
- Conocer las capacidades gráficas de R por medio del paquete *ggplot2.*
- Entender por qué los t-test, anovas ancovas y test no-parametricos son una forma de regresión que es equivalente a un modelo lineal (generalizado o no).
- Entender la estructura de los modelos mixtos desde sus componentes fijos (tamaño de los efectos) y sus componentes aleatorios (descomposición de la varianza).
- Introducir las familias no normales (*binomial, poisson*) y enseñar como trabajar con este tipo de datos por medio de modelos generalizados.
- Evaluar las principales ventajas y desventajas de los procesos de construcción de modelos y de selección.
- Presentar como se realiza el ajuste de los modelos mixtos por técnicas convencionales y bayesianas por medio del paquete *MCMCglmm*.

CONTENIDO RESUMIDO DEL PROGRAMA

UNIDAD No. 1: Estadística básica como ejercicio de manejo de datos en R.

- Aspectos básicos sobre *R*.
- Operaciones básicas en *R*.
- Funciones útiles en *R*.
- Vectores, dataframes y matrices en *R*.
- Estadísticos básicos en *R*.
- Manejo de datos (paquete *dplyr*).
- Ejercicios prácticos en *R*.
- Gracos básicos en *R*.

UNIDAD No. 2: Introducción y repaso sobre modelos lineales

- Introducción a los modelos lineales.
- Interpretación del intercepto, coeficientes de regresión y ajuste de los modelos lineales.
- ANOVAs *vs*. regresión lineal.
- ANCOVAs *vs*. regresión múltiple.
- Interacciones en regresión (entre variables continuas, interacciones cuadráticas y factores) .
- Ajuste e interpretación de regresión múltiple.
- Gráficos avanzados (*e.g., forest plots* y paneles múltiples) por medio del paquete *ggplot2*.

UNIDAD No. 3: Modelos generalizados (GLM) y modelos mixtos (LMM y GLMM)

- Datos binomiales, chi-square vs. regresión logística.
- Conteos, dispersión, over-dispersión y proporciones.
- Construcción y selección de modelos.
- Modelos mixtos lineales.
- Ajuste de modelos mixtos.

- Criterios de selección de modelos (AIC, BIC, QAIC, *stewise selection*).
- Interpretación de modelos mixtos.
- Efectos fijos (coeficientes de regresión) y efectos aleatorios (descomposición de varianza).
- Uso de diferentes paquetes: *nmle*, lme4.

UNIDAD No. 4: Estadística bayesiana y ajuste de modelos vía cadenas de MonteCarlo

- Breve introducción a la estadística bayesiana y a las cadenas de Monte Carlo.
- Inferencia estadística por métodos bayesianos.
- Ajuste de modelos mixtos por técnicas bayesianas.
- Ajuste e interpretación de modelos mixtos con el paquete *MCMCglmm*.
- Uso y selección de *priors*.
- Grafícos y presentación de los modelos (outputs).

UNIDAD No. 5: Inclusión de filogenias en los modelos mixtos

- Estimación de covarianzas filogenéticas.
- Evaluación de las señales filogenéticas.
- Modelo de regresión filogenético.
- Entender la información filogenética como datos jerárquicos.
- Ajuste e interpretación de modelos filogenéticos mixtos con el paquete *MCMCglmm*.
- Tests para evaluar la señal filogenética.

UNIDAD No. 6: Aspectos complementarios de *R*

- El uso de *R Markdown*.
- El uso de *R Sweave*.
- El uso de *R Shiny Web*.
- El uso de *R HTML*.
- El uso de *R Presentation.*

PROPUESTA METODOLÓGICA

Se realizarán clases magistrales y prácticas de clase con una frecuencia semanal. En cada una de las clases se revisarán artículos de caso asociados con las temáticas de clase, pero se espera también poder abordar el análisis de datos de cada uno de los estudiantes del curso para fortalecer sus trabajos de grado. El curso es de carácter teórico-práctico ya que cada sesión se realizará uso de la plataforma *R* y de sus paquetes asociados. Esto último busca que cada estudiante adquiera destrezas con el lenguaje *R*. Por medio de actividades prácticas sobre temas específicos por fuera de clase, los estudiantes tendrán la oportunidad de evaluar diferentes tipos de datos para así mejorar su interpretación sobre los fenómenos biológicos.

CRITERIOS GENERALES DE EVALUACIÓN

- Actividad Práctica No. 01. 10%
- Actividad Práctica No. 02. 10%
- Actividad Práctica No. 03. 10%
- Actividad Práctica No. 04. 10%
- Actividad Práctica No. 05. 10%
- Actividad Práctica No. 06. 10%
- Actividad Práctica No. 07. 10%
- Actividad Práctica No. 08. 10%

- Actividad Práctica No. 09. 10%
- Actividad Práctica No. 10. 10%

BIBLIOGRAFÍA

***Libros de texto*:**

Chang W. 2003. R Graphics CookBook. O'Reilly. 413 p.

Crawley M. 2007. The R Book. John Wiley & Sons. 951 p.

Burnhan K. & D. Anderson. 2002. Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach. Second Edition. Springer. 515 p.

Faraway. J. 2006. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonpaametric Regression Models. Texts in Statistical Science. Taylor & Francis Group. 345 p.

Gelman A. & J. Hill. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Analytical Methods for Social Research. Cambridge. 651 p.

Ives, A. 2018. Mixed and Phylogenetic Models: A conceptual introduction to correlated data. Leanpub. 125 p.

McElreath R. 2016. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. Texts in Statistical Science. Taylor & Francis Group.493 p.

Wickham H. 2013. Ggplot2 Elegant Graphics for Data Analysis. Use R. Srpinger. 268 p.

Zuur A., Ieno, E., Walker N., Saveliev A. & G. Smith. 2009. Mixed Effects Models and Extensions in R. Statistical for Biology and Health. Springer. 580 p.

Revistas de interés:

Methods in ECology and Evolution, Journal of Statistical Software, American Naturalist, Nature, Science, Journal of Animal Ecology, Journal of Evolutionary Biology, Animal Behavior, Behavioral Ecology, Trends in Ecology and Evolution.