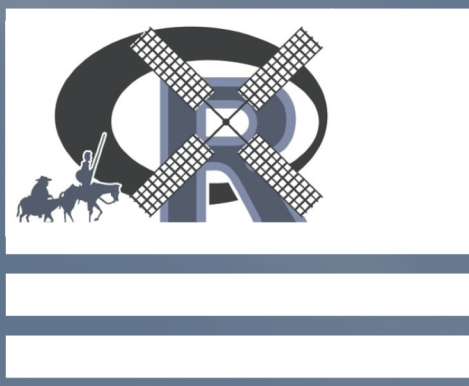
Universidad de Castilla-La Mancha

CAMPUS DE EXCELENCIA INTERNACIONAL

VIII Jornadas usuarios de



Albacete, 17 y 18 de noviembre 2016

http://r-es.org/8jornadasR/

#8JRes

Organizadores:

Comunidad Hispano

DHEP

Departamento de Economía Política y Hacienda Pública, Estadística Económica y Empresarial y Política Económica

DPTO. DE MATEMÁTICAS

EIIAB

Escuela de Ingenieros Industriales de Albacete

Facultad de Ciencias Económicas y Empresariales

www.uclm.es/ab/fcee

Patrocinadores oro:

RStudio

innova-tsn

Microsoft

Editoriales:

CRC Press

Taylor & Francis Group

CHAPMAN & HALL

WILEY

Springer

Bienvenida

En un lugar de La Macha, de cuyo nombre no quiero acordarme... ¡¡celebramos este año las VIII Jornadas de Usuarios de R!!

Estimado/a usuaRio/a,

Te damos la bienvenida a las VIII Jornadas de Usuarios de R organizadas por la Universidad de Castilla-La Mancha. Este congreso continúa con la tradición de proporcionar un punto de encuentro sobre el software estadístico R a nivel nacional. Este año hemos intentado elaborar un programa atractivo que engloba dos charlas invitadas, 5 talleres, 6 sesiones temáticas con presentaciones orales y una sesión de pósters. Desde los Comités encargados de organizar estar jornadas, esperamos que éstas sean productivas y que pases unos días en Albacete.

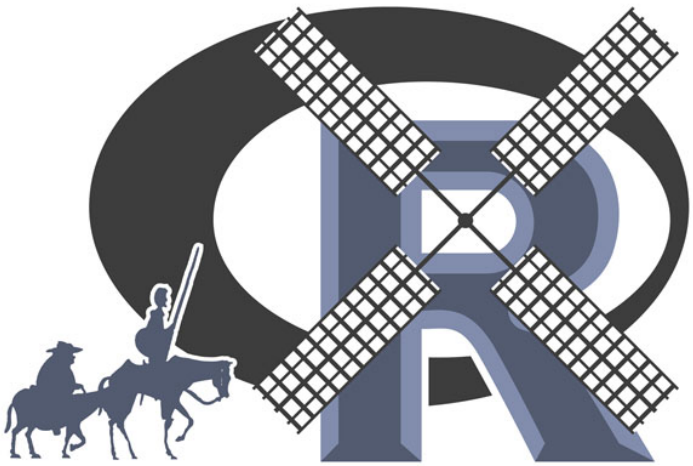
Recibe un cordial saludo.

Comité Organizador Local

Esteban Alfaro-Cortés
José Luis Alfaro-Navarro
María Teresa Alonso-Martínez
Matías Gámez-Martínez
Noelia García-Rubio
Virgilio Gómez-Rubio (Coordinador)
Francisco Palmí Perales
Emilio L. Cano
Francisco Parreño-Torres

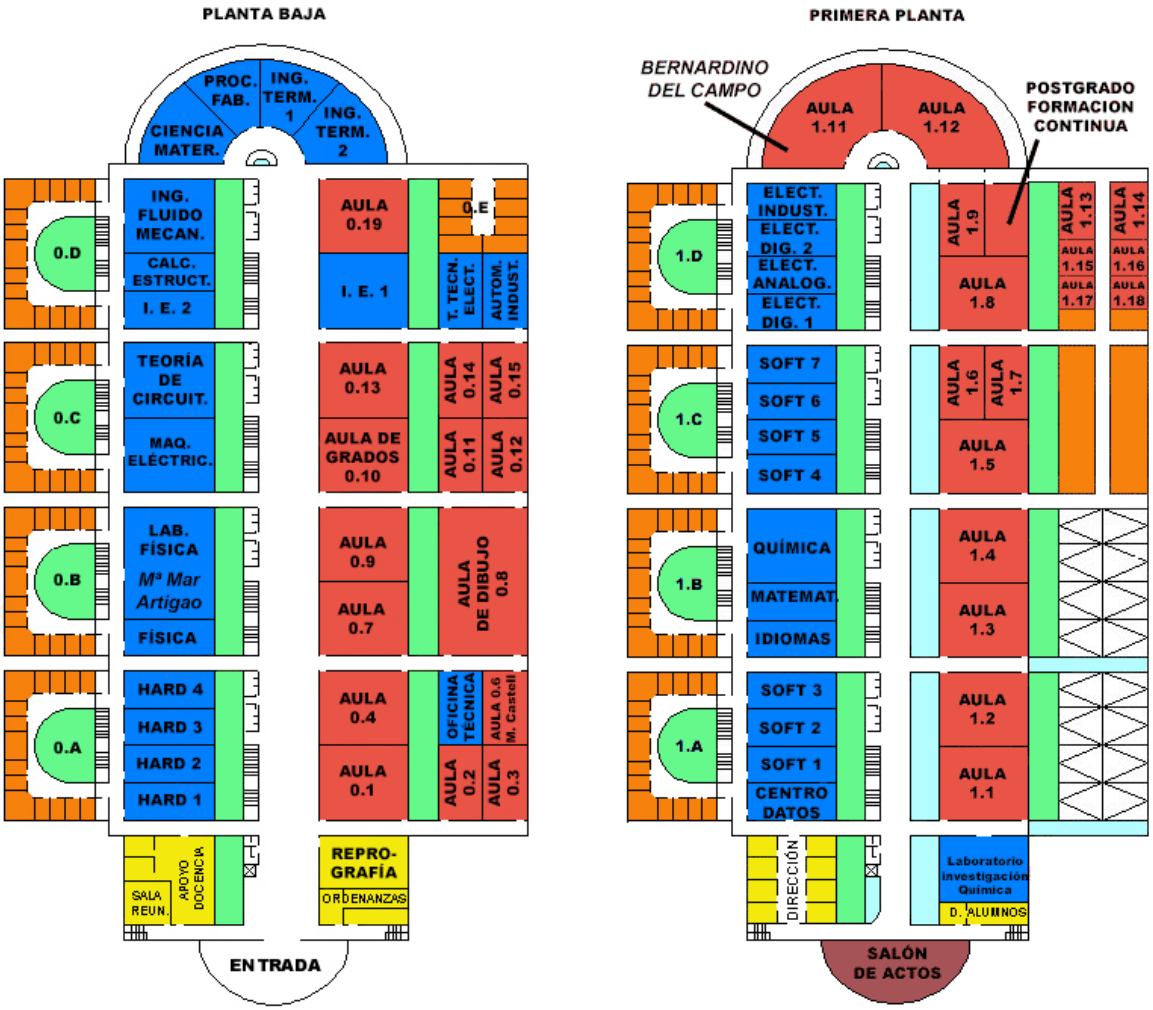
Comité de Programa

Esteban Alfaro Cortés (Coordinador)
José Manuel Benítez
Pedro Concejero Cerezo
Noelia García Rubio (Coordinadora)
Carlos J. Gil Bellosta
Jorge Luis Ojeda Cabrera
Francisco J. Rodríguez Aragón
Miguel Ángel Rodríguez Muíños
Emilio Torres Manzanera



Mapas

Agrupación Politécnica



VIII Jornadas de usuarios de R

Edificios del campus

- Escuela Politécnica Superior
- Facultad de Ciencias Económicas y Empresariales

Programa social

- ClanDestino
- Salones y restaurante POSADA REAL

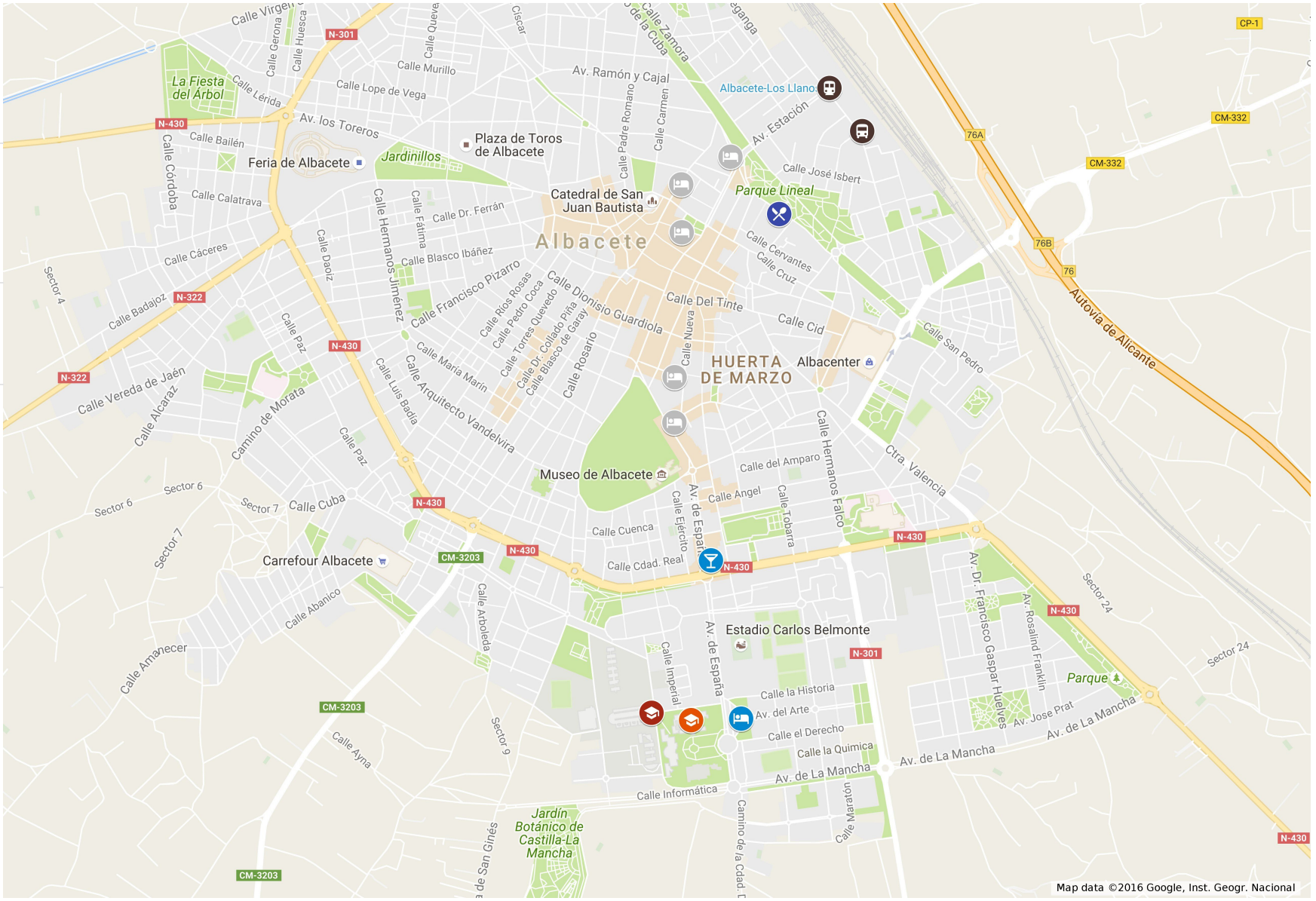
Transporte público

- Albacete-Los Llanos
- Estación de autobuses

Hoteles

- Hotel Universidad
- Hotel Sercotel Los Llanos
- Hotel San Jose Albacete
- Hotel Altozano
- Hotel San Antonio
- Hotel Castilla

Puntos de interés de las VIII Jornadas de usuarios de R (Albacete, noviembre 2016)



Programa

17 de noviembre de 2016

08:30 - 09:00	Entrega de documentación y acreditación	
Salón de Actos		
09:00 - 09:30	Inauguración oficial de las jornadas	
Salón de Actos		
09:30 - 10:30	Comunicaciones I. Bioestadística/Biomedicina/ Ecología	
Aula 0.14	Moderador: Pedro Concejero Cerezo	
	Nucleoplot: Aplicación online basada en R para la visualización del posicionamiento de nucleosomas mediante datos obtenidos por NOME-seq	Francisco Requena Sánchez
	Técnicas de aprendizaje automático aplicadas al análisis de actitudes parentales frente a la vacunación pediátrica	Antonio Maurandi-López, Aurora González Vidal, Alvaro Hernández Vicente, José Antonio Palazón Ferrando, Laura del Río Alonso, Ma Dolores Pérez Cárceles
	Determinación del riesgo microbiológico con R y shiny: el paquete bioinactivation	Alberto Garre Pérez, Pablo S. Fernández Escámez, Jose A. Egea
	Modelización espacio-temporal Bayesiana de distintas enfermedades	Francisco Palmí Perales, Virgilio Gómez Rubio
09:30 - 10:30	Comunicaciones II. Redes sociales y otras aplicaciones de R	
Salón de Actos	Moderador: Esteban Alfaro Cortés	
	Análisis comportamental en redes sociales: Difusión en modelos híbridos	Rafael López, Lourdes Molera, Maria Semitiel, Pedro Noguera
	Multi-dimensional Outlier Detection. An R implementation.	Arturo Azcorra, Luis F. Chiroque, Rubén Cuevas, Antonio Fernández Anta, Henry Laniado, Rose E. Lillo, Juan Romo, Carlo Sguera
	KDD con R.TeMiS, su aplicación en el Proyecto Exhibitium.	José Pino-Díaz, Nuria Rodríguez Ortega, Antonio Cruces Rodríguez, Carmen Tenor Polo, Ana Carmen Benítez Hidalgo, María Casas González, Marieta Jivkova, Carmen Molina, M. Ángel Sánchez Badillo y Bárbara Romero.
	Predicción de destinos en automoción	Carlos Salort Sánchez
	Creación y gestión de un paquete R para análisis de datos demográficos	Pedro J. Pérez, Francisco G. Morillas

10:30 - 11:00	Pausa café	
Salón de Actos		
11:00 - 12:00	Ponencia invitada	
Salón de Actos	rimas: un marco y arquitectura en R para gestión de riesgos en seguridad aérea	David Ríos Insua
12:00 - 13:30	Comunicaciones III. Bioestadística/Biomedicina/ Ecología	
Aula 0.14	Moderador: Virgilio Gómez Rubio	
	Técnicas y paquetes de R para el desarrollo y validación de un modelo predictivo de pérdida de peso inadecuada tras cirugía bariátrica siguiendo las recomendaciones de la declaración TRIPOD	María Elvira Ferre Jaén, Antonio José Fernández López, Antonio Maurandi López
	SAIC50. Aplicación Shiny para el cálculo de la dosis 50	Aurora González-Vidal, Antonio Maurandi-López, Antonia Bernabeu-Esclapez, Antonio J. Perán-Orcajada
	Métodos de puntuación de propensión (propensity score) y técnicas de aprendizaje automático (machine learning) para el equilibrio de covariables en un estudio observacional de comparación de dos técnicas de cirugía bariátrica.	María Elvira Ferre Jaén, Antonio José Fernández López, Antonio Maurandi López
	Una aplicación web interactiva en Shiny para el análisis espacio-temporal de riesgos de mortalidad en áreas pequeñas	A. Adin, J.M. Carrillo, M.D. Ugarte
12:00 - 13:30	Comunicaciones IV. R para los Negocios	
Salón de Actos	Moderador: Francisco J. Rodríguez Aragón	
	Análisis del Abandono en el Sector Financiero	José Miguel Miralles López
	Técnicas de screening y su implementación en R aplicadas en inteligencia de negocio	Pedro Concejero
	Relaciones entre clientes con R	José Manuel Picaza García
	Teoría de cópulas aplicada a la optimización de precios	Pablo Hidalgo García
	Estimación de búsquedas en Google	Andriy Tkachenko
13:30 - 15:30	Comida	
Fac. Económicas		

15:30 - 18:30	Talleres paralelos I (*con café de 17:00 a 17:30*)	
Aula 0.2	Procesamiento de Datos Masivos con SparkR 2.0 (parte 1)	Manuel J. Parra Royón
Aula 0.11	Support Vector Machine: Concepto y Aplicaciones en R	Francisco J. Rodríguez Aragón
Aula 1.14	Análisis de cuestionarios con R. Punto de vista de la consultoría estadística	Antonio Maurandi-López, Álvaro Hernández Vicente, Elvira Ferre Jaén, Antonio José Perán Orcajada, Ana Belen Marín Valverde
18:30 - 19:30	Asamblea Comunidad R-Hispano	
Salón de Actos		
19:30 - 21:30	Visita guiada a Albacete	
Salida desde la	*entrada de la Agrupación Politécnica*	
21:30 - 24:00	Cena social en el Restaurante Posada Real (*C/ Alcalde Conangla 18*)	

18 de noviembre de 2016

09:00-10:00	Ponencia invitada	
Salón de Actos	Presente y futuro del diseño óptimo de experimentos. Una perspectiva computacional	Jesús Fernando López Fidalgo
10:00-11:00	Sesión de póster y café	
Salón de Actos	Un modelo de precios dinámicos basado en Regresión Isotónica	J. Santos Domínguez-Menchero, Emilio Torres-Manzanera
	ShinyEST: una aplicación interactiva para el autoaprendizaje de la Estadística	Daniel Gómez, María Dolores Molina, Julio Mulero, María José Nueda, Aurora Pascual
	Estadística con R en el Grado en Administración y Dirección de Empresas	Lourdes Molera Peris, Fuensanta Arnaldos García, Ma Teresa Díaz Delfa, Úrsula Faura Martínez, Isabel Parra Frutos, Juan José Pérez Castejón
	Enviromental data analysis with R	Carmen Capilla
	Una aplicación web para el análisis univariante y multivariante de datos metabolómicos	Ibon Martínez-Arranz, Itziar Mincholé, Maite Gutiérrez-Calzada, Cristina Alonso
	Detección de outliers mediante secuencias no monótonas	Victor Mariscal, Victoria López, Diego Urgelés
	Integración de datos medioambientales desde portales Open Data con herramientas R	Pavel LLamocca y Victoria López
11:00-14:00	Talleres paralelos II	
Aula 0.2	Procesamiento de Datos Masivos con SparkR 2.0 (parte 2)	Manuel J. Parra Royón
Software 5	Estimación en Áreas Pequeñas	Virgilio Gómez Rubio
Hardware 2	knitr y RMarkdown para la Generación Automática de Informes	Jorge Luis Ojeda Cabrera

14:00 - 16:00	Comida	
Fac. Económicas		
16:00 - 17:15	Comunicaciones V: Docencia y otras aplicaciones de R	
Salón de Actos	Moderadora: Noelia García Rubio	
	Enseñanza de fundamentos de la investigación y análisis de datos en Tercer Ciclo. Evaluación de resultados de aprendizaje	Laura del Río Alonso, María Francisca Carreño Fructuoso, Elvira Ferre Jaén, Aurora González Vidal, Antonio Maurandi López, Álvaro Hernández Vicente, Fernando Pérez Sanz, Jose Antonio Palazón Ferrando
	R en Estadística Aplicada en la Universidad de Alcalá	Marcos Marvá, Fernando San Segundo
	Gestión académica de una escuela universitaria con R y shiny	Oscar Perpiñán Lamigueiro
	Generación de Información Sintética en el ámbito demográfico-actuarial	Francisco G. Morillas Jurado, Pedro J. Pérez Vázquez
	Evaluación de hipótesis en la construcción de tablas de vida. Una representación gráfica	Francisco Morillas, Jose M. Pavía y Josep Lledó
16:00 - 17:15	Comunicaciones VI: Paquetes e Interfaces	
Aula 0.11	Moderador: Emilio López Cano	
	clickR: Un paquete para facilitar la redacción de informes	Victoria Fornés Ferrer, David Hervás Marín
	Interfaz de Usuario Shiny para una función: Prototipado rápido de Shiny apps	Jorge Luis Ojeda Cabrera
	R-package. Distribución Marshal-Olkin Zipf (MOEZipf)	Ariel Duarte-López, Aina Casellas Torrentó, Marta Pérez- Casany
	exreport: Un paquete de R para el análisis reproducible de datos experimentales	Jacinto Arias, Javier Cózar
	Microsoft R server	Juan Carlos Rodríguez García
17:15 - 17:30	Pausa café	
Salón de Actos		
17:30 - 18:00	Entrega del premio al mejor trabajo joven y clausura de las jornadas	
Salón de Actos		

Charlas invitadas

RIMAS: un marco y arquitectura en R para gestión de riesgos en seguridad aérea

David Ríos Insua

Instituto de Ciencias Matemáticas (ICMAT)

Como requiere la OACI, los países deben desarrollar los denominados Programas de Seguridad Estatal para promover una aproximación proactiva a la gestión de la seguridad aérea a nivel estatal. Con ellos se apoya la toma de decisiones estratégicas y la asignación de recursos a las áreas de aviación de mayor riesgo. Como las herramientas disponibles, basadas en matrices de riesgos, eran de escaso rigor, AESA, RAC e ICMAT has desarrollado una metodología nueva basada en herramientas y modelos de Ciencias de la Decisión y de Datos, implementada en R denominada RIMAS. Esto ha permitido a AESA mejorar el apoyo a sus procesos de toma de decisiones y alcanzar considerables ahorros, como se explicará en la presentación.

Presente y futuro del diseño óptimo de experimentos. Una perspectiva computacional

Jesús López Fidalgo

Universidad de Navarra

El diseño óptimo de experimentos es una disciplina que va mas allá del diseño clásico de experimentos, aunque de alguna manera lo incluye. Tiene la ventaja de abordar todas aquellas herramientas de la estadística que utilizan datos experimentales, entendidos estos en un sentido muy amplio. Tradicionalmente ha recibido muchas críticas por el empeño excesivo en optimizar la experimentación suponiendo un modelo determinado a priori, antes de tener los datos. Lo cierto es que éste no es un problema menor, pero esta teoría pone su punto de mira en descubrir donde se puede conseguir la mayor información y eso es siempre un valor añadido. Otra de las críticas habituales es la dificultad para calcular diseños óptimos. Por eso el desarrollo de algoritmos eficientes es algo, más que importante, estrictamente necesario. Herramientas como R son de una relevancia esencial en este sentido. Una tercera crítica, más reciente, se dirige al hecho de que el diseño óptimo busca minimizar el número de experimentos a realizar, cuando hoy día la preocupación parece estar más bien en qué hacer con las grandes cantidades de datos que circulan a nuestro alrededor. También en esto el diseño experimental puede ayudar a detectar dónde está la información escondida en una maraña intratable de datos. Estas ideas describen básicamente el contenido que tendrá la charla, con la inclusión de algunos ejemplos reales.

Talleres

Support Vector Machine: Concepto y Aplicaciones en R

Francisco J. Rodríguez Aragón

INNOVA - TSN

1. Introducción: La botella de Klein
2. Concepto de Support Vector Machine (SVM)
3. Supestos de los modelos tipo SVM
4. Tipos de Support Vector Machines
5. Un ejemplo clásico: La implementación de la función XOR
6. Predicción y fiabilidad de resultados de los SVM
7. Extensiones y generalizacions de los SVM
8. Ejercicio de SVM y ejemplos de aplicaciones de SVM con R

Bibliografía

▷ Carmona Suárez J. (2014) Tutorial sobre Máquinas de Vectores Soporte (SVM) UNED http://www.ia.uned.es/~ejcarmona/publicaciones/[2013-Carmona]%20SVM.pdf

▷ Cortes, C.; Vapnik, V. (1995) Support-vector networks. Machine Learning, 20(3), 273-297

▷ Gareth, J.; Witten, D.; Hastie, T. y Tibshirani R. (2013) An Introduction to Statistical Learning with Applications in R Springer Science + Business Media New York ISBN 978-1-4614-7137-0

Análisis de cuestionarios con R. Punto de vista de la consultoría estadística

Antonio Maurandi-López, Álvaro Hernández Vicente, Elvira Ferre Jaén, Antonio José Perán Orcajada, Ana Belen Marín Valverde

Univeridad de Murcia

En un servicio central, como el Servicio de Apoyo a la Investigación de la Universidad de Murcia, hemos de analizar con frecuencia conjuntos de muy diversos ámbitos que requieren un análisis muy similar, es el caso de los cuestionarios o encuestas. Con este tipo de datos es usual encontrarse con preguntas o ítems tipo Likert, ordinales o variables medidas en intervalos, agrupados por temáticas o bloques.

En nuestro servicio hemos llegado a pseudo-automatizar este tipo de análisis, partiendo de los paquetes 'Rmarkdown' y 'knitr' asistidos por otros paquetes como 'likert', 'pander', 'xtable', 'ztable', 'tables', 'DT' y algunas funciones definidas ad hoc de forma que obtenemos unos resultados muy vistosos, que sientan la base para decidir futuros análisis a aplicar.

En este taller se analizaría un conjunto de datos desde el principio, lectura de datos, explicando las estrategias seguidas y las diversas opciones que surgen en cada momento y cómo resolverlas, para llegar finalmente a un informe en formato 'pdf' o 'html'.

Estimación en Áreas Pequeñas

Virgilio Gómez Rubio

Universidad de Castilla-La Mancha

Este tutorial se centrará en distintos métodos para el análisis de datos provenientes de encuestas para producir estimaciones en áreas pequeñas y/o subdominios. La metodología que se cubrirá en el tutorial incluye estimadores basados en el diseño del muestreo, y estimadores basados en modelos (incluyendo métodos Bayesianos y basados en la verosimilitud). Asimismo, toda la metodología presentada irá acompañada de numerosos ejemplos desarrollados con R. Los paquetes utilizados serán principalmente de la 'Task View' sobre 'Official Statistics'.

La estructura del tutorial es:

1. Short introduction to survey sampling strategies with R: simple random sampling, systematic sampling, clustered sampling, two-stage sampling
2. Design based-estimators: Horvitz-Thompson, generalized regression (GREG) and calibration estimators
3. Model-based estimators: Fay-Herriott estimator, linear regression for area and unit level models
4. Synthetic and composite estimators
5. Mixed-effects models: area and unit level models with random effects, EBLUP estimators
6. Models with spatial random effects: spatial EBLUP for area and unit level models
7. Bayesian inference for Small Area Estimation: area and unit level models
8. Non-linear models: disease mapping, estimation of unemployment

Los materiales del tutorial se encuentran disponibles en GitHub.

knitr y RMarkdown para la Generación Automática de Informes

Jorge Luis Ojeda Cabrera

U. de Zaragoza

Taller Markdown básico:

- Metodología: ejemplos y conceptos básicos
- Requisitos: Conocimientos básicos de R, RStudio y una instalación RStudio.
- Tópicos:

  * Sintaxis básica de Markdown. RStudio como herramienta markdown. Generación de documentos preprocesado y resultados intermedios.

  * R Chunks: opciones básicas gráficos, salida de markdown, otros tipos de chunks.

Generación de Informes con markdown:

- Metodología: ejemplos y conceptos básicos
- Requisitos: Conocimientos básicos de R, RStudio, Rmarkdown y una instalación RStudio.
- Tópicos:

  * Paquetes RMarkdown y knitr estructura básica.

  * Utilidades knitr y RMarkdown para el preprocesado de ficheros markdown. Relación con Shiny, y otros paquetes útiles.

Big Data con R

Manuel J. Parra Royón

Universidad de Granada

1. Introducción.
    - ▷ R (Intro breve, todo el mundo lo conoce)
    - ▷ Spark y el análisis de datos masivos
    - ▷ Estructura y modelo de computación
    - ▷ Comparación con otros paradigmasVentajas y desventajas
    - ▷ ¿Por qué Spark + R?
    - ▷ Evolución del proyecto
    - ▷ Arquitectura
    - ▷ Ventajas
    - ▷ Limitaciones
    - ▷ Novedades en la v2.0
2. Puesta en marcha de la infraestructura SparkR 2.0
    - ▷ Preparación del entorno base Inicialización de MVs
    - ▷ Puesta en marcha de los servicios necesarios
    - ▷ Validación de la inicialización
    - ▷ Ejecución de sistema de Notebooks Jupyter
    - ▷ Muy breve introducción al entorno Jupyter
3. SparkR: Introducción y preparación del entorno inicial
    - ▷ Validación del las variables de entorno y Paths
    - ▷ Biblioteca SparkR
    - ▷ Inicialización de SparkR:Herramientas de gestión: Jobs, etc.
    - ▷ Creación de un notebook básico con dataset.
4. SparkR: Dataframes y API
    - ▷ Trabajo con HDFS, Dataframe, etc.
    - ▷ Funciones de la API de Spark
    - ▷ Trabajo con fuentes de datos
    - ▷ Funciones estadísticas y matemáticas
5. SparkR: Algoritmos Machine Learning
    - ▷ GLM, NB, KMeans, ...
    - ▷ Biblioteca MLLIB
6. Visualización de datos
7. Conclusiones
8. Material

Trabajos candidatos al Premio Joven

Sesión	Título	Autores
Comunicaciones I	Determinación del riesgo microbiológico con R y shiny: el paquete bioinactivation	Alberto Garre Pérez, Pablo S. Fernández Escámez, Jose A. Egea
Comunicaciones I	Modelización espacio-temporal Bayesiana de distintas enfermedades	Francisco Palmí Perales, Virgilio Gómez Rubio
Comunicaciones I	Nucleoplot: Aplicación online basada en R para la visualización del posicionamiento de nucleosomas mediante datos obtenidos por NOME-seq	Francisco Requena Sánchez
Comunicaciones II	Multi-dimensional Outlier Detection. An R implementation	Arturo Azcorra, Luis F. Chiroque, Rubén Cuevas, Antonio Fernández Anta, Henry Laniado, Rose E. Lillo, Juan Romo, Carlo Sguera
Comunicaciones III	Una aplicación web interactiva en Shiny para el análisis espacio-temporal de riesgos de mortalidad en áreas pequeñas	A. Adin, J.M. Carrillo, M.D. Ugarte
Comunicaciones IV	Teoría de cópulas aplicada a la optimización de precios	Pablo Hidalgo García
Comunicaciones VI	exreport: Un paquete de R para el análisis reproducible de datos experimentales	Jacinto Arias, Javier Cózar

Enseñanza de fundamentos de la investigación y análisis de datos en Tercer Ciclo. Evaluación de resultados de aprendizaje

Laura del Río Alonso, María Francisca Carreño Fructuoso, Elvira Ferre Jaén, Aurora González Vidal, Antonio Maurandi López, Álvaro Hernández Vicente, Fernando Pérez Sanz, Jose Antonio Palazón Ferrando.

Universidad de Murcia

En el marco de la formación de doctorado y dentro del programa Escuela Internacional de Doctorado de la Universidad de Murcia (curso 2015-2016), se han desarrollado una serie de cursos bajo el título general de "Diseño de experimentos y fundamentos de análisis de datos". Con 140 inscripciones, los asistentes a estos cursos pertenecen a titulaciones de muchas áreas de conocimiento, con perfiles y niveles muy heterogéneos y en general, limitada o nula experiencia en manejo de datos, programación o R.

Nos planteamos en esta comunicación una revisión de los resultados y posibles ventajas de la metodología empleada.

Estos cursos se han impartido de forma secuencial, incrementando la complejidad de la materia en los distintos cursos pudiendo los alumnos participar en ellos de forma independiente. Se inician con un curso básico de R y Rstudio, y a continuación uno sobre investigación reproducible y creación de documentos científicos.

La metodología empleada ha sido eminentemente práctica, combinando las actividades presenciales con el trabajo personal de cada alumno. Para el seguimiento de la actividad se recurre al aula virtual de la Universidad de Murcia, que nos permite realizar un seguimiento individualizado de la participación y esfuerzo de cada alumno, monitorizando los resultados de su aprendizaje.

Se ha evaluado el grado de satisfacción de los alumnos con la metodología de trabajo y la percepción subjetiva de las competencias adquiridas mediante una encuesta al finalizar el proceso formativo.

Presentaremos los resultados del análisis de adquisición de competencias y habilidades por parte de los alumnos, así como la reflexión y evaluación de la metodología utilizada para el uso de R como instrumento muy relevante en la enseñanza universitaria.

Generación de Información Sintética en el ámbito demográfico-actuarial

Francisco G. Morillas Jurado, Pedro J. Pérez Vázquez

Universitat de València

En el estudio de ciertos fenómenos la generación de números aleatorios es utilizada para reproducir el comportamiento de variables de interés, ya sea por la imposibilidad de obtener datos reales, por el coste económico o por el coste temporal. En el ámbito demográfico-actuarial, y en particular en el estudio de la supervivencia y la mortalidad -donde el experimento es difícil de replicar- el uso de técnicas de simulación que ayuden a generar escenarios sintéticos de mortalidad es muy importante. Los escenarios generados pueden ser utilizados en: la construcción de tablas de mortalidad sintéticas; en la obtención de estadísticos de resumen; en la construcción de intervalos de confianza; en la generación de representaciones gráficas que ayuden a analizar el fenómeno de manera conveniente; en realizar de forma más sencilla análisis por edad o por grupo de edad; en la obtención de escenarios extremos;?Así mismo, los datos generados pueden servir como datos de entrada en procesos más complejos, que no tienen por qué ser lineales: la tarificación, el aprovisionamiento, el cálculo de valores de seguridad que ayuden a garantizar la solvencia de la compañía o de un proceso, son sólo algunos de los ejemplos que podemos encontrar en el ámbito actuarial y financiero. Así, en este trabajo se presenta un ejemplo de simulación numérica esencial que es utilizado tanto en investigación como en docencia de master, en las asignaturas Análisis Demográfico no paramétrico y Procesos Estocásticos, donde entre otros objetivos se encuentra el de fomentar el conocimiento y uso de R en el ámbito de la simulación numérica, aplicado en demografía general o actuarial.

R en Estadística Aplicada en la Universidad de Alcalá

Marcos Marvá, Fernando San Segundo

Universidad de Alcalá

El objetivo de esta comunicación es ilustrar el uso de R como herramienta docente en Estadística Aplicada a otras disciplinas. En concreto, queremos presentar nuestra experiencia en la Universidad de Alcalá en la que, desde el año 2011, hemos usado R en los Grados de Biología y Biología Sanitaria. Nuestro enfoque se basa en incorporar la programación como base fundamental de la enseñanza de la Estadística. El hilo conductor de toda la asignatura es la programación con R, para la construcción de simulaciones y modelos. A esto añadimos herramientas manipulativas, como GeoGebra, o de cálculo, como Calc o WolframAlpha. Con la intención de dar autonomía a quien quiera adentrarse en la estadística, el proyecto consta de tres partes: un libro para la parte teórica, unos tutoriales para los aspectos computacionales y un conjunto suficientemente rico de cuestionarios aleatorios para la (auto)evaluación, todo ello accesible libremente en http://www.postdata-statistics.com.

Aunque, como decimos, ese material se ha puesto a prueba en grados de Biología, nos hemos esforzado en no escribir un curso de "Bioestadística con R". Que, sin duda, tendría interés. Pero nuestro objetivo era adoptar un punto de vista más amplio y tratar de escribir un curso de Introducción a la Estadística de tipo conceptual, que pueda ser utilizado tanto para el autoestudio como para un conjunto muy amplio de primeros cursos de Estadística en grados universitarios. La parte computacional queda cubierta por una colección de tutoriales, uno por capítulo del libro, en los que se utiliza R de forma protagonista (pero también GeoGebra, Calc o Wolframalpha) para abordar los aspectos prácticos y computacionales del curso. Y, por último, pero no menos importante, otra componente de nuestro trabajo es el proyecto que hemos denominado ExamineR.

ExamineR se basa en el paquete exams de R, creado en la Facultad de Economía de la Universidad de Viena (ver referencia más abajo). Ese paquete permite generar de forma automática y aleatorizada cuestionarios y exámenes con distintos tipos de preguntas (verdadero/falso, respuesta múltiple, respuesta numérica, tipo cloze, etc). Los cuestionarios pueden crearse como documentos pdf, html o en formato XMl para su utilización en plataformas educativas tipo Moodle. A pesar la potencia de esa herramienta, una limitación es la falta de bancos de preguntas disponibles para quieres quieran usarlo. Nuestro proyecto incluye un repositorio en GitHub (https://github.com/PostDataStatistics/ExamineR) de preguntas de Estadística, que cubre todos los aspectos básicos de un curso de Introducción a la Estadística. Las preguntas se han diseñado para que sea fácil su adaptación a distintos idiomas y están inicialmente disponibles en inglés y español. Actualmente estamos trabajando en una aplicación web que permitirá el acceso a las preguntas del repositorio y la generación de cuestionarios a usuarios sin conocimientos de R; incluida la generación de un examen equivalente para cada alumno, junto con sus correspondientes soluciones.

Paquete exams de R: Bettina Gruen, Achim Zeileis (2009). Automatic Generation of Exams in R. Journal of Statistical Software 29(10), 1-14. URL http://www.jstatsoft.org/v29/i10/.

Gestión académica de una escuela universitaria con R y shiny

Oscar Perpiñán Lamigueiro

Universidad Politécnica de Madrid

La gestión académica de una Escuela Universitaria abarca tareas dispares, como la elaboración y publicación de horarios de docencia y tutorías por profesor, calendarios académicos, o el análisis de tasas de resultados por asignatura y titulación. En el desempeño de mi cargo como Jefe de Estudios en la Escuela Técnica Superior de Ingeniería y Diseño Industrial (UPM), una Escuela con más de 2700 alumnos y 150 profesores, he desarrollado durante los últimos meses un conjunto de herramientas basadas en R y shiny para facilitar y automatizar estas tareas. En esta presentación resumiré los aspectos técnicos más destacados de estas herramientas. También aportaré las lecciones que he aprendido acerca del desarrollo de interfaces de usuario y aplicaciones web, y su empleo por un conjunto de usuarios con habilidades diversas.

Evaluación de hipótesis en la construcción de tablas de vida. Una representación gráfica

Francisco Morillas, Jose M. Pavía y Josep Lledó

Universitat de València

En desarrollo informático observado en los últimos tiempos ha permitido incorporar nueva información en las técnicas estadísticas utilizadas en el campo actuarial y demográfico. Además, el continuo incremento de la esperanza de vida ha intensificado los estudios relacionados sobre la metodología y datos utilizados en la construcción de las tablas de mortalidad.

Generalmente, las tablas de mortalidad utilizadas por organismos públicos y privados son construidas con datos agregados y asumen hipótesis implícitas como (i) distribución uniforme de los fallecidos y (ii) sistema demográfico cerrado. La eliminación de ambas hipótesis con datos reales requiere su evaluación previa mediante test espaciales y funcionales tales como el CLF test, el Maximum Absolute Deviation test, el Kolgomorov-Smirnov test y el test Geometric que se encuentran recogidos en los packages spatstat y GoFKernel.

En muchas ocasiones, la dificultad en la representación gráfica de los resultados de los distintos p.values no permite expresar los datos con suficiente claridad para la comprensión del lector. En este trabajo se muestra como visualizar de manera clara y sintética los diferentes p-values de una batería de test de hipótesis (o de un mismo test con diferente nivel de significatividad) utilizando el package raster.

Análisis del Abandono en el Sector Financieero

José Miguel Miralles López

Grupo Cooperativo Cajamar

Analizaremos el abandono de cliente en las entidades financieras definiendo el tipo de abandono, pasando por las distintas fases de pre-procesacimiento de datos y finalmente determinaremos el mejor modelo que se adapta a predecir el abandono del cliente. Finalmente, realizaremos una segmentación de valor con el fin de poder tener una visión mas completa de la estrategia de negocio.

Técnicas de screening y su implementación en R aplicadas en inteligencia de negocio

Pedro Concejero

Telefónica Investigación y Desarrollo

Las metodologías conocidas como 'screening' o detección temprana son ampliamente utilizadas en medicina especialmente en áreas como el cáncer. Aunque la variedad de técnicas médicas es impresionante, una característica común de muchas de ellas es el objetivo de identificar marcadores con buena capacidad predictiva y que se consigan mediante pruebas baratas o rápidas, aunque luego requieran la realización de pruebas diagnósticas más costosas. Como tales no son herramientas diagnósticas, y por tanto pueden tener un porcentaje significativo de fallos (falsos positivos y negativos), pero deben tener la gran ventaja de su rapidez, facilidad de uso y bajo coste. Desde el punto de vista metodológico el gran reto es calibrar adecuadamente la capacidad predictiva y la repercusión de los posibles fallos, así como incorporar el coste en el proceso.

Esta estrategia es perfectamente aplicable en entornos de inteligencia de negocio, en los que uno de los retos es identificar, cuanto antes mejor, individuos que respondan a un perfil determinado: que estén en riesgo de abandonar, que sean favorables a ampliar los servicios de los que disponen, entre otros muchos. Desde el punto de vista metodológico el objetivo es similar al de detección temprana. Y también lo son los retos, aunque por supuesto sin las implicaciones sobre la salud que sí existen en medicina.

Esta propuesta pretende extender, prototipar con un conjunto de datos disponible públicamente, ajustar y demostrar estas técnicas con las librerías actualmente disponibles de R (ROCR, pROC, hmeasure). Plantearemos funciones para el cálculo de AUC (o también Wilcoxon-Mann-Whitney) de forma masiva y mejoras de la visualización de las curvas ROC asociadas, así como el cálculo de valores predictivos positivos y negativos incorporando información de la prevalencia en el proceso. Además, probaremos e incorporaremos, en la medida de lo posible, los procedimientos roccurve, comproc, rocreg y predcurve propuestos por una investigadora clave en este campo, Margaret Pepe (http://research.fhcrc.org/diagnostic-biomarkers-center/en/software/rocbasic.html).

Relaciones entre Clientes con R

José Manuel Picaza García

BBVA

Como parte del proceso de innovación activa del BBVA, desde el área de Riesgos de España y junto con BBVA Data & Analytics, en el Dpto. de Modelos y Herramientas estamos desarrollando aplicaciones para entender mejor las relaciones que existen de nuestros clientes entre ellos y con otros particulares y empresas de fuera del banco. Para llevar a cabo estas investigaciones hemos desarrollado proyectos de gestión de datos para representar a los clientes como un problema de grafos. Hemos trabajado las librerías de representación visual de grafos (visNetwork) que R está desarrollando activamente y las librerías estadísticas sobre grafos que implementa (igraph) En el marco de estas jornadas nos gustaría presentar un prototipo desarrollado por nosotros este año, que sirvió de base para los desarrollos estratégicos que se están implementando. Es interesante que podáis asomaros un poco al tipo de trabajo que se desarrollan en estas áreas de entidades financieras, donde tradicionalmente no había mucha innovación, pero en las que cada día es más importante el valor añadido que aportan la I+D, así como el rol que herramientas como R, Python, neo4j y las plataformas de Big Data juegan en esta evolución. La base del trabajo es diseñar un grafo en el que los nodos van a ser los clientes, pegando la información relevante para el seguimiento de su salud financiera, y crear tantas aristas como tipos de relaciones diferentes que puedan existir entre ellos. De entre las más interesantes son las que relacionan productos bancarios, transacciones y la información pública de empresas. Una vez entendida y estructurada esta información, hemos desarrollado funciones para pintar las relaciones de la manera que mejor nos permitiera visualizar las redes de clientes que surgen a partir de un CIF/NIF concreto. Estamos trabajando en procesos que evalúen estos grafos para asociar a cada cliente una puntuación que tenga en cuenta la red en la que está embebido el mismo.

Teoría de Cóputlas aplicada a la optimización de precios

Pablo Hidalgo García

INNOVA-TSN

Un problema habitual en el sector retail es el de conseguir fijar el precio de un producto que permita maximizar el beneficio. En este caso, se propone abordar el problema a través de la teoría de cópulas. La teoría de cópulas intenta buscar la función conjunta que mejor relacione (copule) dos funciones de distribución marginales. Esta teoría, con un amplio recorrido en el sector económico, se ha utilizado poco fuera de ese ámbito. Hemos aplicado la teoría de cópulas con éxito al sector retail permitiendo relacionar las ventas de un producto con su precio para proponer el precio esperado que maximice las ganancias. Además, con ayuda de una cópula n-dimensional, se pueden incluir efectos de productos canibalizadores que afecten a las ventas del producto en estudio y así obtener relaciones entre los distintos productos que conforman el catálogo. Una vez ajustado el modelo, se optimiza el problema de forma que se proponga el precio óptimo para el conjunto de productos que componen el catálogo haciendo uso del algoritmo de búsqueda local Hill Climbing.

Estimación de búsquedas en Google

Andriy Tkachenko

Havas Media

Se trata de un proyecto realizado con R y aplicado al mundo de la publicidad. Básicamente es un análisis de detección de tendencias, estacionalidades y el impacto de la inversión publicitaria sobre las búsquedas en Google de cerca de 350 anunciantes, mediante modelos de regresión. En la ponencia se presentará todos los pasos del proyecto, desde extracción hasta las estimaciones, junto con los resultados más destacados.

Análisis comportamental en redes sociales: Difusión en modelos híbridos

Rafael López, Lourdes Molera, Maria Semitiel, Pedro Noguera

Universidad de Murcia

En la literatura del análisis de redes sociales, se asume la teoría de la conexión preferencial como la más acertada para representar procesos de incorporación de nuevos nodos a la red. Sin embargo, desde nuestra perspectiva, el modelo híbrido constituye una alternativa que mejora a la ofrecida por la teoría de la conexión preferencial. Entre sus ventajas se puede destacar que contempla un gran abanico de posibles distribuciones del grado de la red. En este trabajo se presta una especial atención a los procesos que implican la difusión de innovaciones. Existe numerosa literatura que hace referencia a la importancia de la distribución del grado de una red dentro de los procesos de difusión, donde cabría pensar que si se modifica la red, por ejemplo, cambiando la función de distribución que sigue el grado de los nodos, se producirían modificaciones en los resultados conocidos sobre la difusión de la innovación.

Si se aplica la regla de la conexión preferencial, los nuevos nodos se unirán a los ya existentes con mayor grado, por lo tanto si la difusión se produce por contacto, son los nodos de mayor grado los que mayor probabilidad tienen de ser innovados dentro del proceso de difusión. Los nuevos nodos tendrán una probabilidad muy alta de ser innovados, ya que los hubs (nodos más centrales o con más conexiones dentro de la red) son más propensos a ser innovados y, en consecuencia, los nodos que se unen a estos. Al incorporar una componente aleatoria para la unión de nuevos nodos, se espera crear nuevas conexiones entre grupos que no se relacionan o que tienen caminos más largos para comunicarse en el caso puramente preferencial, lo que propiciará un proceso de difusión mucho mayor y más rápido.

En el análisis de este proceso es necesario encontrar una herramienta que mida la influencia de cada nodo sobre el proceso de difusión de innovaciones. La variable o criterio puede ser diverso, como la posición del nodo en la red, la relación del nodo con sus vecinos, o el número de conexiones que tiene cada nodo, entre otras muchas. Además, cada una de esas herramientas podría usarse en la elección de los nodos por los que comenzar el proceso de difusión para acelerarlo. Ha habido recientes aportaciones a la literatura, como la propuesta del índice de centralidad de comunicación o del índice de centralidad de difusión. En este trabajo se presenta una nueva medida que mejora las predicciones que aportan los dos índices anteriores, en la difusión de una innovación dentro de una red.

Multi-dimensional Outlier Detection. An R implementation.

Arturo Azcorra, Luis F. Chiroque, Rubén Cuevas, Antonio Fernández Anta, Henry Laniado, Rose E. Lillo, Juan Romo, Carlo Sguera

IMDEA Networks Institute

Con el auge de las redes sociales online, la cantidad de datos disponibles para los sociólogos ha aumentado considerablemente en los últimos años. Sin embargo, la mayoría de los métodos tradicionalmente usados para procesar y extraer información no suelen escalar para procesar millones de usuarios de las redes sociales digitales.

En este trabajo, proponemos un método escalable, no supervisado, y de propósito general, para identificar muestras atípicas basado en análisis de datos funcionales. Este método encaja perfectamente para ser usados con datos de redes sociales online. Además de la escalabilidad, el método propuesto es capaz de identificar atípicos de 3 distintos tipos: forma, amplitud y magnitud, aportando una clasificación más valiosa.

Hemos sido capaces de evaluar nuestro método usando un conjunto de datos de 5.6 millones de usuarios de Google+ con 23 variables cada usuario. Además, los usuarios atípicos detectados por nuestro método tienen aspectos interesantes.

La charla constará de una descripción formal de nuestro método, así como una detallada explicación de su implementación en R.

KDD con R.TeMiS, su aplicación en el Proyecto Exhibitium.

José Pino-Díaz, Nuria Rodríguez Ortega, Antonio Cruces Rodríguez, Carmen Tenor Polo, Ana Carmen Benítez Hidalgo, María Casas González, Marieta Jivkova, Carmen Molina, M. Ángel Sánchez Badillo y Bárbara Romero

Universidad de Málaga

El proyecto "Generación de conocimiento sobre exposiciones artísticas temporales para su reutilización y aprovechamiento multivalente"(Exhibitium), presentado por el grupo de investigación iArtHis_Lab de la UMA, ha sido premiado en la convocatoria 2014 de premios de investigación de la Fundación BBVA, en la categoría de Humanidades Digitales. La base de datos de Exhibitium ha sido analizada mediante R.TeMiS [R Text Mining Solution] con la finalidad de generar nuevo conocimiento sobre las .[ex]posiciones artísticas temporales.[en] su condición de fenómenos culturales complejos y como elementos estratégicos en la generación de dinámicas sociales y movimientos económicos.

Se presenta el paquete R.TeMiS, su aplicación y los resultados obtenidos en el análisis de datos (minería de textos) del Proyecto Exhibitium.

Predicción de destinos en automoción

Carlos Salort Sánchez

Accenture

Utilizando los datos obtenidos a través del puerto OBD-II de los coches modernos, en este trabajo predecimos el destino al que se dirige el usuario en semi tiempo real, mediante algoritmos de machine learning.

Creación y gestión de un paquete R para análisis de datos demográficos

Pedro J. Pérez, Francisco G. Morillas

Universitat de València

El sistema de paquetes de R que permite ampliar las funcionalidades de R base ha sido una de las claves de su éxito, de forma que actualmente, CRAN alberga más de 8000 paquetes. Para poder alojar un paquete en CRAN, este debe cumplir unos requisitos formales y de calidad que se detallan en el manual "Writing R extensions"del R Core Team.

Disponer de un paquete de uso privado dentro de un grupo, ya sea grupo de investigación o docente, es también de gran utilidad pues posibilita el agrupar un conjunto de funciones, datos y documentación requerida para el desarrollo de un proyecto concreto o para la enseñanza de una asignatura. Si el paquete está además alojado en un servicio como Github, los participantes en el proyecto o curso tendrán agrupados y disponibles todos los recursos necesarios en un solo paquete.

En la presentación se ilustrará el proceso de creación y gestión de un paquete de R para una asignatura de gestión de cartera de seguros con datos demográficos. Durante la presentación se presentará de forma sencilla lo siguiente: creación de un paquete de R con devtools, incorporación de documentación con roxygen2, cómo alojar el paquete en Github, modificación del paquete añadiendo nuevas funciones y cómo actualizar el paquete en Github utilizando Git.

Nucleoplot: Aplicación online basada en R para la visualización del posicionamiento de nucleosomas mediante datos obtenidos por NOME-seq.

Francisco Requena Sánchez

Universidad de Granada. Centro de Genómica e Investigación Oncológica (GENyO)

La epigenética se ha convertido en unos de los campos con mayor progreso de los últimos años. A partir de una misma secuencia de ADN existen más de 200 tipos celulares en el cuerpo humano

Terapias innovadoras para enfermos de cáncer basadas en la epigenética ya están en el mercado, conocer cómo a través de una misma secuencia de ADN inalterable para todas las células de nuestro cuerpo, da lugar a tal variabilidad de células, es un objetivo de la investigación biomédica.

La técnica NOME-seq desarrollada en el año 2012 por el equipo de TK. Kelly permite la obtención de información de dos pilares fundamentales de la epigenética como es la metilación en las hebras de ADN y el posicionamiento de nucleosomas. La técnica permite conocer en una misma cadena de ADN secuenciada, ambos elementos de forma simultánea, lo que abre un abanico de posibilidades al poder estudiar como afectan estos dos factores en, por ejemplo, sitios promotores de genes ante ambientes o estados diferentes de la célula.

El problema actual, reside en la dificultad de obtener y analizar la información que arroja la técnica una vez secuenciadas las muestras. No existe actualmente ninguna herramienta que permita un ágil tratamiento de los datos y la obtención de resultados de forma automática.

A continuación presentamos Nucleoplot, una "webtoolçonstruída completamente en R que mediante el input de secuencias tratadas por NOME-seq, el usuario puede analizar sus datos en cuestión de segundos. Algunas claves del proyectos:

1. La aplicación web se ha implementado gracias a Shiny, usando sus funciones tanto para la interfaz web (ui.R) como para el mecanismo interno de procesamiento de datos (server.R).

2. Debido a la naturaleza del problema, se implementó una imagen reactiva en la interfaz, que con modificaciones en el panel de control, el usuario puede ver como varía la imagen (1). Una vez que el usuario esté conforme con el resultado obtenido, el programa permite descargar los resultados finales:

2.1 Tablas de resultados en formato .csv para que el usuario pueda trabajar con los datos desde Excel.

2.2 Imagen vectorial de las gráficas lollipop (1). La ventaja de la imagen vectorial (.svg) frente a una .png, reside en la facilidad de la modificación posterior del usuario de cualquier elemento, ya sea tamaño, color...lo cual facilita la publicación de la imagen en la revista científica.

2.3 Gracias a la implementación de RMarkdown, el programa arroja un informe en formato html, con información relevante acerca de como ha analizado las muestras, los datos introducidos, que parámetros resultantes ha indicado el usuario y en definitiva de todo el proceso ejecutado por el programa, evitando así que se convierta en una caja negra para el usuario. El informe creado a través de Rmarkdown, es totalmente personalizado, todos sus elementos son creados de forma única para cada usuario, en función de los datos y parámetros escogidos.

En definitiva, Nucleoplot integra R junto al potencial de Rmarkdown y Shiny, implementando una solución única para la comunidad científica. Los próximos pasos del proyecto es terminar de perfilar la interfaz web y su funcionamiento en un servidor y la creación del paquete de R, nucleoplot en Bioconductor, para que cualquier usuario pueda tener acceso al código fuente.

Me gustaría exponer este trabajo a la comunidad de usuarios de R y poder enseñar el gran potencial de este lenguage usando las diferentes herramientas (Shiny, Rmarkdown) que tenemos a nuestra disposición, para conseguir, en definitiva, una solución integral a problemas biológicos.

Técnicas de aprendizaje automático aplicadas al análisis de actitudes parentales frente a la vacunación pediátrica

Antonio Maurandi-López, Aurora González Vidal, Alvaro Hernández Vicente, José Antonio Palazón Ferrando, Laura del Río Alonso, Mª Dolores Pérez Cárceles

Servicio de Apoyo a la Investigación (SAI), Universidad de Murcia

Nos planteamos el análisis de diversos perfiles de actitudes y conocimientos de las madres y padres con hijos menores de 14 años frente a las vacunas pediátricas en la Región de Murcia.

Partiendo de los datos recogidos en un cuestionario (n=1119), se construyeron grupos de encuestados mediante técnicas de aprendizaje no supervisado: análisis de correspondencias y análisis de cluster (k-means y jeráquico). En una fase posterior se caracterizaron los grupos mediante gráficos de perfiles y análisis de varianza.

Se han identificado varios determinantes de la reticencia de madres y padres hacia la vacunación de sus hijos en la población murciana. Este trabajo no sólo contribuye al diagnóstico de perfiles parentales reticentes hacia la vacunación", sino que, además, aporta un procedimiento de análisis con técnicas estadísticas multivariantes que puede ser de gran utilidad en el análisis de conjuntos de datos donde existan múltiples interacciones.Nos planteamos el análisis de diversos perfiles de actitudes y conocimientos de las madres y padres con hijos menores de 14 años frente a las vacunas pediátricas en la Región de Murcia.

Partiendo de los datos recogidos en un cuestionario (n=1119), se construyeron grupos de encuestados mediante técnicas de aprendizaje no supervisado: análisis de correspondencias y análisis de cluster (k-means y jeráquico). En una fase posterior se caracterizaron los grupos mediante gráficos de perfiles y análisis de varianza.

Se han identificado varios determinantes de la reticencia de madres y padres hacia la vacunación de sus hijos en la población murciana. Este trabajo no sólo contribuye al diagnóstico de perfiles parentales reticentes hacia la vacunación", sino que, además, aporta un procedimiento de análisis con técnicas estadísticas multivariantes que puede ser de gran utilidad en el análisis de conjuntos de datos donde existan múltiples interacciones.

Determinación del riesgo microbiológico con R y shiny: el paquete bioinactivation

Alberto Garre Pérez, Pablo S. Fernández Escámez, Jose A. Egea

Universidad Politécnica de Cartagena

El sistema de paquetes que ofrece el lenguaje de programación R permite extenderlo con funcionalidades con las que no contaba inicialmente. Además, el paquete shiny permite crear aplicaciones web que son capaces de utilizar funciones de R. Es decir, permite a los usuarios utilizar las funcionalidades de R sin que sea necesario conocimientos de programación. Esto expande enormemente el rango de usuarios potenciales de R.

En este trabajo se presenta una extensión de R dirigida a la microbiología predictiva: el paquete bioinactivation (Garre et al., 2016). La microbiología predictiva es la rama de la microbiología que intenta describir la evolución de una población microbiana bajo unas condiciones ambientales dadas variables con el tiempo. Una de las aplicaciones de esta ciencia es el estudio de la seguridad microbiológica de los alimentos, ya que permite describir la evolución de los microorganismos patógenos durante el ciclo de vida del producto. Es decir, a través de las diversas fases de inactivación y esterilización que atraviesa el producto desde que se recolectan las materias primas hasta el momento de su consumo.

Los modelos matemáticos para la descripción de la respuesta de una población microbiana bajo condiciones ambientales dinámicas más usuales en industria e investigación forman un sistema de ecuaciones diferenciales y algebraicas. Tanto la generación de predicciones como el ajuste de este tipo de modelos a datos experimentales a datos experimentales son tareas complejas que requieren el uso de técnicas numéricas. No existen actualmente en el mercado herramientas capaces de resolver este problema, así que los grupos de investigación deben de recurrir a software in-house. El paquete bioinactivation pretende saciar esta necesidad, incluyendo funciones para el cálculo de predicciones (tanto en forma de curvas como intervalos de predicción) como para el ajuste de curvas a datos experimentales. Para ello, utiliza las funciones para la resolución numérica de sistemas de ecuaciones diferenciales implementadas en el paquete deSolve (Soetaert et al., 2010), y estimación de parámetros e inferencia Bayesiana incluídas en el paquete FME (Soetaert y Petzoldt, 2010). Con el propósito de hacer esta herramienta lo más accesible posible se ha puesto a disposición del público por medio de CRAN y se ha desarrollado una aplicación de shiny que incluye las funciones del paquete más comunes.

Bibliografía

Alberto Garre, Pablo S. Fernandez and Jose A. Egea (2016). bioinactivation: Simulation of Dynamic Microbial Inactivation. R package version 1.1.2. http://CRAN.R-project.org/package=bioinactivation.

Karline Soetaert and Thomas Petzoldt (2010). Inverse Modelling, Sensitivity and Monte Carlo Analysis in R Using Package FME. Journal of Statistical Software, 33(3), 1-28. URL http://www.jstatsoft.org/v33/i03/.

Karline Soetaert, Thomas Petzoldt, R. Woodrow Setzer (2010). Solving Differential Equations in R: Package deSolve Journal of Statistical Software, 33(9), 1–25. URL http://www.jstatsoft.org/v33/i09/.

Técnicas y paquetes de R para el desarrollo y validación de un modelo predictivo de pérdida de peso inadecuada tras cirugía bariátrica siguiendo las recomendaciones de la declaración TRIPOD

María Elvira Ferre Jaén, Antonio José Fernández López, Antonio Maurandi López

Servicio de Apoyo a la Investigación (SAI), Universidad de Murcia

Un grupo internacional de investigadores y estadísticos publicaron, en el año 2015, la Declaración TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) con la finalidad de ayudar a mejorar la calidad de los informes de estudios que desarrollan, validan o actualizan modelos de predicción clínica. La guía TRIPOD comprende una lista de comprobación de 22 ítems que han sido elegidos porque la evidencia empírica indica que esta información es esencial para entender el estudio, valorar la fiabilidad y la relevancia de sus hallazgos y, por otra parte, la ausencia de esta información está asociada con estimaciones sesgadas.

Partiendo de una cohorte formada por 300 pacientes obesos mórbidos intervenidos mediante cirugía bariátrica, se realizó un estudio, ajustándose a las recomendaciones TRIPOD, con el objetivo de desarrollar y validar un modelo predictivo de pérdida de peso inadecuada al año de la cirugía bariátrica.

La selección del mejor modelo predictivo de regresión logística multivariante se realizó a partir del cálculo de todas las posibles ecuaciones de regresión ("best subset") que se podían obtener con las variables predictoras seleccionadas. Se utilizó como criterio principal de selección la optimización del índice de Akaike (AIC), y se siguió la regla de 10-15 eventos por variable predictora para evitar la sobresaturación (overfitting) del modelo.

Para la evaluación del rendimiento del modelo se emplearon las medidas habituales de rendimiento global, calibración, discriminación junto con medidas específicas de utilidad clínica. En cuanto a la validación interna del modelo final respecto a los parámetros de calibración y discriminación, se utilizaron técnicas de remuestreo (bootstrapping) y penalización de los coeficientes de regresión (shrinkage) previa cuantificación del exceso de optimismo y estimación del sesgo de dichos coeficientes. Por último, y para la presentación gráfica del modelo final en un formato práctico, se elaboró un Nomograma.

En el desarrollo y validación del modelo predictivo de pérdida de peso inadecuada al año de la cirugía bariátrica, se han seguido las recomendaciones de la declaración TRIPOD, y se han utilizado diversas técnicas y paquetes de R.

Modelización espacio-temporal Bayesiana de distintas enfermedades.

Francisco Palmí Perales, Virgilio Gómez Rubio.

Universidad de Castilla la Mancha

En las últimas décadas, se han desarrollado distintos modelos dentro de la creación de mapas de enfermedades y la detección de clusters. Estos modelos tienen como objetivo observar y analizar el comportamiento espacial (o espacio temporal) de cualquier fenómeno dentro de una región de estudio, poniendo especial atención en aquellas localizaciones/regiones a los que correspondan valores por encima de lo esperado. En este trabajo hemos obtenido un modelo Bayesiano espacio-temporal basado en el modelo de Besag, York y Mollié (BYM) y desarrollado a partir de la combinación de las ideas propuestas por Abella et al en [1] y por Downing et al en [2], con el objetivo de analizar conjuntamente la muerte de cuatro cánceres: cáncer de esófago, cáncer de ovarios, cáncer de vejiga y cáncer de pulmón. La parte computacional la hemos realizado utilizando R y R2Winbugs. En los resultados podemos observar el comportamiento espacial y temporal de las cuatro enfermedades global y específicamente. El paso inmediato a este trabajo es la aplicación del modelo a nivel municipal, el cual, eleva considerablemente el tiempo computacional de ejecución del modelo. Actualmente se está trabajando en la combinación de MCMC e INLA con el objetivo de reducir dicho tiempo computacional.

Bibliografía

Abellan J., Richardson S. and Best N. (2008). Use of Space-Time to Investigate the Stability of Patterns of Disease. Enviromental Health Perspectives 116, 1111-1119.

Downing A., Forman D., Gilthorpe M., L Edwards K. and OM Manda S. (2008). Joint disease mapping using six cancers in the Yorkshire region of England. International Journal of Health Geographics 7:41.

SAIC50. Aplicación Shiny para el cálculo de la dosis 50

Aurora González-Vidal, Antonio Maurandi-López, Antonia Bernabeu-Esclapez, Antonio J. Perán-Orcajada

Universidad de Murcia

En el contexto de experiencias de investigación usuales en ciencias experimentales y biomédicas es común la necesidad de calcular el índice ÏC50", entre otros valores de dosis 50 (DL50, DI50, CE50), para una sustancia determinada, esto es, la capacidad inhibitoria media de la sustancia con respecto al proceso biológico (crecimiento, supervivencia, producción de una proteína, . . . etc) del cultivo en concreto. Dicho índice indica la concentración de sustancia necesaria para inhibir el proceso biológico en un 50 % (a la mitad).

En la mayoría de los experimentos, la curva dosis-respuesta tiene una forma sigmoidea lo que implica que el cálculo del IC50 no sea simple para el investigador.

Es por ello que se ha desarrollado una aplicación con acceso web de sencillo manejo que facilita al investigador el cálculo de este parámetro empleando el paquete Shiny de R, permitiendo la carga de datos de forma transparente y la elección del método, de entre varios, para calcular el IC50. La aplicación tiene implementados 5 métodos actualmente, documentados en la bibliografía científica para el cálculo del IC50: regresión lineal, regresión lineal con transformación en el eje x, regresión lineal con transformación en ambos ejes, regresión lineal eliminando puntos y función logística de los 4 parámetros.

En esta presentación se mostrará dicha aplicación.

Métodos de puntuación de propensión (propensity score) y técnicas de aprendizaje automático (machine learning) para el equilibrio de covariables en un estudio observacional de comparación de dos técnicas de cirugía bariátrica

María Elvira Ferre Jaén, Antonio José Fernández López, Antonio Maurandi López

Servicio de Apoyo a la Investigación (SAI), Universidad de Murcia

Existen estudios observacionales en los que la asignación de los pacientes a los distintos tratamientos no es aleatoria, los grupos de tratamiento que se comparan difieren a menudo en covariables importantes que pueden estar relacionadas con las variables de respuesta, y que por tanto, pueden conducir a estimaciones sesgadas de los tratamientos comparados. Los estudios observacionales basados en puntuaciones de propensión pueden corregir este sesgo dando lugar a estimaciones insesgadas de los efectos del tratamiento.

Partiendo de un estudio de cohorte formado por 300 pacientes obesos mórbidos intervenidos de cirugía bariátrica se realiza un estudio comparativo de dos técnicas quirúrgicas: Bypass Gástrico Laparoscópico (BGL) y Gastrectomía Vertical Laparoscópica (GVL). Dado que los datos recogidos proceden de un diseño observacional,las comparaciones entre los grupos pueden verse afectadas por sesgos tanto de confusión como de selección de los pacientes.

Para garantizar la compatibilidad de los grupos analizados se llevó a cabo un análisis de propensión ("Propensity Score Analysis") con el objetivo de reconstruir un hipotético proceso de aleatorización que habría producido unos grupos de tratamiento como los obtenidos tras el emparejamiento ("matching") por las puntuaciones de propensión. Se obtuvieron así dos grupos de pacientes (uno operado mediante BGL y otro mediante GVL) balanceados en más 22 variables sociodemográficas y clínicas.

Para el cálculo de las puntuaciones de propensión se emplearon (y compararon) las siguientes técnicas de aprendizaje automático ("machine learning"): support vector machine (SVM), random forest (RF), baggin, boosting, gradient boosting, decisión tree (algoritmo C5.0). de arboles de decisión. También se empleo una regresión logística multivariante por pasos ("backward").

Se realizó un emparejamiento de 2 BGL por cada 1 GVL, mediante el método de vecino más próximo con un caliper de 0,2, con el paquete 'nonrandom'. El método support vector matching (SVM) fue el que más variables consiguió balancear (21 variables de un total de 22), y consiguiendo emparejar a 154 pacientes (5 pacientes tuvieron un matching incompleto). Para la medición del equilibrio entre las covariables se empleo la diferencia de medias estandarizadas antes y después del matching, y esta no fue mayor de 0.20 en 21 variables. También se comprobó que el equilibrio entre las ratios de varianza de cada covariable, entre los grupos comparados, fuera próxima a 1.

Una vez seleccionados los pacientes de cada grupo se procedió a realizar diversas comparaciones sobre la efectividad y las complicaciones de cada una de estas dos técnicas quirúrgicas.

Todos los análisis realizados (propensity score analysis, machine learning, generalized lineal models, etc) se llevaron a cabo utilizado diversas técnicas y paquetes de R.

Una aplicación web interactiva en Shiny para el análisis espacio-temporal de riesgos de mortalidad en áreas pequeñas

A. Adin, J.M. Carrillo, M.D. Ugarte

Universidad Pública de Navarra

La representación cartográfica de enfermedades (en inglés, "disease mapping"), es un área de investigación de creciente interés en epidemiología y salud pública. En la actualidad ya no se representan sólo los casos de mortalidad o incidencia en forma de mapas, sino que suelen representarse riesgos o tasas suavizados a partir de modelos estadísticos cada vez más sofisticados. Esto es debido a la gran variabilidad que muestran las medidas clásicas de estimación de riesgos (como la razón de mortalidad estandarizada o las tasas brutas) cuando se analizan enfermedades raras o cuando se calculan sobre áreas pequeñas.

El objetivo de este trabajo es crear una herramienta sencilla que permita a cualquier usuario, en particular a los investigadores de centros de salud y otros organismos públicos, aplicar modelos estadísticos para el análisis espacio-temporal de riesgos o tasas de mortalidad o incidencia. Para ello presentaremos una aplicación web utilizando "Shiny". Además de las medidas clásicas para el análisis descriptivo de los riesgos (o tasas) de mortalidad/incidencia para datos espacio-temporales, se han implementado distintos modelos [1] que permiten suavizar estos riesgos utilizando la metodología INLA (integrated nested Laplace approximations) recientemente propuesta por Rue et al. (2009) [2]. En la implementación de los modelos se ha prestado especial atención a aspectos de identificación de los modelos y a la posibilidad de considerar distintos tipos de interacciones espacio-temporales, cuestiones que pueden resultar complejas para usuarios no expertos que se enfrenten a la modelización en R-INLA.

Bibliografía

M.D. Ugarte, A. Adin, T. Goicoa, A.F. Militino. (2014). On fitting spatio-temporal disease mapping models using approximate Bayesian inference, Statistical Methods in Medical Research, 23:507-530.

H. Rue, S. Martino, and N. Chopin. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, Journal of the Royal Statistical Society, Series B, 71:319-392.

clickR: Un paquete para facilitar la redacción de informes

Victoria Fornés Ferrer, David Hervás Marín

IIS La Fe

En esta presentación se mostrarán las funciones del paquete en desarrollo clickR, cuyo objetivo es facilitar la generación de informes de calidad. El paquete consta de funciones para la generación de descriptivos en formato de tabla lista para publicar, gráficos tanto descriptivos como de resultados y otras herramientas útiles como la función mine.plot() que permite explorar la base de datos de manera rápida y visual. También se incluyen funciones para la generación de tablas de resultados de todo tipo de modelos (lineal, glm, gam, cox, etc.).

El paquete está integrado con knitr y markdown, por lo que permite aprovechar todas sus ventajas.

Interfaz de Usuario Shiny para una función: Prototipado rápido de Shiny apps.

Jorge Luis Ojeda Cabrera

Universidad de Zaragoza

Una App no es más que una funcionalidad, es decir una función, a la que se acopla un interfaz de usuario cuyo objetivo es la introducción de los datos necesarios para el cálculo y la presentación de los resultados del mismo. Bajo la premisa de que, como analistas o científicos de datos, nuestra principal labor es el desarrollo de técnicas, es decir de funciones, que permitan una diversidad de análisis y cálculos orientados a este fin, el objetivo de este trabajo es presentar un marco que permitan recrear una interfaz de usuario basada en Shiny para una función, de modo que el foco del trabajo del analista o científico de datos sea principalmente la funcionalidad que desear presentar. A este fin se hace uso de las capacidades funcionales de R para abordar la creación de un Interfaz de Ususario Web para una función utilizando el paquete Shiny.

Entre otras posibilidades, este I.U. para funciones se puede utilizar para el testeo rápido de la funcionalidad, validación de la entrada de datos, desarrollo de análisis y cálculos en directo, ejemplos para enseñanza, desarrollo rápido de nuevas funcionalidades (de Apps, de addings para RStudio, etc...), ...

R-package. Distribución Marshal-Olkin Zipf (MOEZipf)

Ariel Duarte-López, Aina Casellas Torrentó, Marta Pérez- Casany

Universidad Politécnica de Cataluña.

En los últimos años hemos podido observar como ha ido ganando espacio, dentro de la comunidad científica, todo lo relacionado con el análisis de grandes volúmenes de datos. Contar con herramientas que permitan una rápida comprensión de los datos facilita la investigación. Una de estas herramientas especialmente útil en redes sociales son los grafos. Dado un grafo, uno de los aspectos básicos es el estimar la distribución de probabilidad asociada al grado de sus nodos.

En varios artículos se ha presentado la distribución Zipf como una buena candidata a la hora de ajustar datos provenientes de diferentes áreas del conocimiento como, por ejemplo, linguística, biomedicina, redes sociales, etc. Sin embargo, aunque la distribución Zipf goce de esta popularidad, en algunos casos el ajuste obtenido es muy mejorable puesto que se trata de una distribución con un solo parámetro. Para mejorar dicho ajuste proponemos utilizar la distribución MOEZipf(1). Una generalización biparamétrica de la Zipf mucho más flexible en los primeros valores del dominio de la distribución.

El principal objetivo de este trabajo es proveer a la comunidad científica la implementación de un paquete de R que permita utilizar de forma sencilla la distribución MOEZipf. El paquete de R permite el cálculo de probabilidades, percentiles, funciones acumuladas, así como la generación de números aleatorios que sigan esta distribución. Por otra parte, también permite calcular el estimador verosímil dado un conjunto de datos, así como el íntervalo de confianza de los parámetros estimados. Todas las funcionalidades son presentadas a través de ejemplos que utilizan datos reales, ilustrando las mejoras obtenidas mediante la nueva distribución.

(1) https://arxiv.org/pdf/1304.4540.pdf

exreport: Un paquete de R para el análisis reproducible de datos experimentales

Jacinto Arias, Javier Cózar

Universidad de Castilla-La Mancha

exreport es un paquete de R concebido con el fin de proporcionar una herramienta robusta para el análisis de datos experimentales. Su principal objetivo es proporcionar de manera transparente reproducibilidad y automatización a la hora de analizar y preparar nuestros resultados durante el proceso de investigación y la elaboración de publicaciones.

Éste paquete permite describir los datos mediante un modelo general sobre el que es posible definir transformaciones, calcular diversas métricas y aplicar contrastes de hipótesis para finalmente obtener representaciones en forma tabular o gráfica.

Una vez diseñado el análisis es posible generar informes de manera automática, en formato HTML o PDF, con el fin de visualizar la información o acceder de manera sencilla al código LaTeX o las imágenes con el fin de incluirlas directamente en una publicación.

El paquete está disponible en CRAN y en su ultima versión de desarrollo en la página http://exreport.jarias.es, donde pueden encontrarse numerosos ejemplos así como la documentación.

MICROSOFT R SERVER

JUAN CARLOS RODRIGUEZ GARCIA

MICROSOFT

Cómo integrar modelos R en entornos corporativos de alto rendimiento, incluyendo aspectos tales como la distribución en clusters/bases de datos, el despliegue de tareas de scoring en infraestructura cloud y el streaming de datos para evitar las limitaciones de memoria

Posters

Un modelo de precios dinámicos basado en Regresión Isotónica

J. Santos Domínguez-Menchero, Emilio Torres-Manzanera

Universidad de Oviedo

Los precios de las habitaciones de hotel, así como el número de habitaciones disponibles a través de los canales de venta online, varían día a día. Es natural preguntarse cuándo un turista debería reservar una habitación para conseguir un precio más bajo, o cuándo los gestores del hotel deberían supervisar el precio de sus competidores, variar sus ofertas, o buscar otros canales de venta.

En este trabajo se analizan los precios ofertados a través de los más importantes canales de distribución en Internet por hoteles en ciudades españolas con diferentes tipologías turísticas, en un período de ocho meses. Se plantea un modelo de precios dinámicos mediante técnicas de Regresión Isotónica, implementando nuevos algoritmos mediante software totalmente desarrollado en R.

El modelo explica el ahorro que le supondría al consumidor comprar con l días de antelación y consigue detectar estrategias útiles tanto para el comprador como para los gestores del hotel.Los precios de las habitaciones de hotel, así como el número de habitaciones disponibles a través de los canales de venta online, varían día a día. Es natural preguntarse cuándo un turista debería reservar una habitación para conseguir un precio más bajo, o cuándo los gestores del hotel deberían supervisar el precio de sus competidores, variar sus ofertas, o buscar otros canales de venta.

En este trabajo se analizan los precios ofertados a través de los más importantes canales de distribución en Internet por hoteles en ciudades españolas con diferentes tipologías turísticas, en un período de ocho meses. Se plantea un modelo de precios dinámicos mediante técnicas de Regresión Isotónica, implementando nuevos algoritmos mediante software totalmente desarrollado en R.

El modelo explica el ahorro que le supondría al consumidor comprar con días de antelación y consigue detectar estrategias útiles tanto para el comprador como para los gestores del hotel.

ShinyEST: una aplicación interactiva para el autoaprendizaje de la Estadística

Daniel Gómez, María Dolores Molina, Julio Mulero, María José Nueda, Aurora Pascual

Universidad de Alicante

La Estadística es un área de gran importancia en los estudios de Ciencias Sociales y, en ocasiones, presenta grandes dificultades de aprendizaje por parte de los estudiantes de titulaciones no técnicas. Conscientes del atractivo que suponen las nuevas tecnologías para las nuevas generaciones hemos desarrollado una serie de recursos docentes que el alumno podrá utilizar de forma interactiva desde su móvil, tablet u ordenador que persiguen la mejora del proceso enseñanza-aprendizaje.

En particular, hemos diseñado una aplicación denominada ShinyEST que, a partir de datos generados aleatoriamente o bien a partir de datos introducidos por el usuario, proporciona las soluciones al alumno para que éste compruebe sus resultados. ShinyEST ha sido creada con el paquete denominado Shiny del software estadístico R que permite la creación de aplicaciones web interactivas. En este trabajo, presentamos unos recursos docentes diseñados con Shiny que ponen al alcance de los alumnos tantos ejercicios como deseen, permitiéndoles entrenar sus capacidades matemáticas y estadísticas de manera individual desde su propia casa.

Estadística con R en el Grado en Administración y Dirección de Empresas

Lourdes Molera Peris, Fuensanta Arnaldos García, Mª Teresa Díaz Delfa, Úrsula Faura Martínez, Isabel Parra Frutos, Juan José Pérez Castejón

Universidad de Murcia

Muchos aspectos son los que han influido a la hora de tomar la decisión de abordar las prácticas de Inferencia Estadística en el Grado en Administración y Dirección de Empresas de la Universidad de Murcia con el software R. El hecho de que se trate de un programa de libre acceso, muy extendido en la comunidad científica, tanto en el ámbito docente como en el investigador, su versatilidad y el desarrollo de interfaces más sencillas para su uso han sido factores decisivos, pero también el cambio en la orientación del proceso de enseñanza-aprendizaje y la necesidad de poder tratar eficientemente grandes cantidades de datos cada vez más accesibles.

En particular, con R se han podido trabajar, de forma relativamente sencilla, nociones y técnicas que resultaban muy rígidas en software estadístico tradicional o excesivamente tediosas con una hoja de cálculo. De este modo, la facilidad para la obtención de gráficos y resultados ha permitido centrar más la atención en la interpretación y en los posibles efectos de cambios de escenario.

Durante las clases se ha utilizado RStudio, y se han elaborado ficheros html con RMarkdown para facilitar al alumno el uso de R y favorecer su autonomía en las sesiones. También se han realizado transparencias y materiales de autoevaluación con Slidify, así como aplicaciones en Shiny para visualizar datos y reforzar conceptos que suelen ser difíciles de entender. En general, la experiencia ha sido bastante favorable, aun teniendo en cuenta el coste de entrada en el uso del software y el perfil de los alumnos de titulaciones de ciencias sociales.

Enviromental data analysis with R

Carmen Capilla

Universidad Politécnica de Valencia

This paper presents the analysis of enviromental observations using R. Air quality models are estimated , and their performance is evaluated to forecast pollutants concentrations in an urban area. Non parametric tests are applied to study trends and seasonal cycles of carbon monoxide levels. The multivariate assessment of trends of water quality observations, is performed using non-parametric tests. The covariance inversion test supported rejection of the hypothesis of no trend in the variables defined with each combination of water quality parameters and month. There is heterogeneity between the trends in the different combinations and an overall trend is not representative. The partial Mann-Kendall is employed to analyze the trends of each physicochemical variable in the study months. The multivariate techniques that have been used, by utilizing the information in the correlation structure of the data set, provide efficient changes detection for quality assurance and more understandable data analysis outputs for decision-making.

Una aplicación web para el análisis univariante y multivariante de datos metabolómicos

Ibon Martínez-Arranz, Itziar Mincholé, Maite Gutiérrez-Calzada, Cristina Alonso

OWL Metabolomics

La investigación metabolómica ha evolucionado considerablemente, sobre todo en la última década. En el transcurso de esta evolución, el interés por las disciplinas 'omicas' es ahora más evidente que nunca. Sin embargo, el futuro de la metabolómica dependerá de su capacidad para encontrar una interpretación biológica de los resultados. Por esa razón, la minería de datos y el estudio de las rutas metabólicas constituyen una tarea central en el flujo de trabajo de la metabolómica, por lo que se requiere de un conocimiento detallado de bioinformática y de un software especializado. Además, las ciencias 'ómicas' tienden a trabajar conjuntamente para ofrecer distintos puntos de vista de un mismo fenómeno, por lo que la difusión y comparación de resultados debe ser fluida y eficaz.

En este contexto se ha desarrollado OWL Stat App, una aplicación web implementada en R y basada en el concepto de programación reactiva del paquete Shiny, combinado con desarrollo en HTML, para compartir, comparar y analizar datos metabolómicos. Esta aplicación combina un potente análisis de datos univariantes y multivariantes con herramientas de rutas metabólicas y de visualización para facilitar la interpretación de los resultados.

Las pruebas estadísticas univariantes y herramientas de visualización disponibles incluyen un heatmap y un gráfico volcán interactivo que permite seleccionar un metabolito de interés y se proporciona una descripción de dicho compuesto de acuerdo con las bases de datos de HMDB (Human MetabolomeDatabase) y KEGG (KyotoEncyclopedia of Genes and Genomes), así como un diagrama de cajas, histogramas, gráficos de densidad y gráficos Q-Q, análisis de normalidad mediante la prueba Shapiro-Wilk, prueba t de Student, análisis de valores atípicos, ratios y análisis ROC. Todas las pruebas estadísticas mencionadas y salidas gráficas pueden volver a calcularse aplicando diferentes transformaciones predefinidas. OWL Stat App también incluye los siguientes análisis multivariantes: análisis de componentes principales (ACP); estudio de correlación entre muestras; mapas de calor. Todos los cálculos se realizan con el paquete caret de clasificación y el paquete ROCR para visualizar los rendimientos de los distintos metabolitos como clasificadores. El paquete pheatmap se utiliza para la elaboración de mapas de calor. Finalmente, para ayudar en la interpretación de los cambios del metabolito en un contexto biológicamente significativo, esta aplicación combina las herramientas analíticas basadas en R con la identificación de los metabolitosen las diferentes rutas metabólicas en las que intervienen, incluyendolas bibliotecas de rutas metabólicas SMPDB (Small MoleculePathwayDatabase) y propias desarrolladas en nuestro laboratorio.

Detección de outliers mediante secuencias no monótonas

Victor Mariscal, Victoria López, Diego Urgelés

Universidad Complutense de Madrid

En este trabajo se muestra la implementación con R de un algoritmo para detección de outliers basado en cadenas no monótonas y búsqueda de patrones con el algoritmo KMP. Se muestran también los resultados comparados con otros algoritmos tradicionales y se prueba la validez del método.

Integración de datos medioambientales desde portales Open Data con herramientas R

Pavel LLamocca y Victoria López

Universidad Complutense de Madrid

En la actualidad, muchos gobiernos están publicando miles de conjuntos de datos. Como consecuencia, la cantidad de aplicaciones basadas en Open Data está incrementándose. Sin embargo, cada gobierno tiene sus propios procedimientos para publicar y se produce una gran variedad de formatos sin existir un estándar internacional. En este trabajo se presenta una solución capaz de integrar datos medioambientales. La solución además permite al usuario visualizar y hacer análisis sobre los datos en tiempo real. Una vez que el proceso de integración está realizado, todos los datos de cada gobierno poseen el mismo formato y se pueden lanzar procesos de análisis de una manera más computacional.

El trabajo tiene tres partes fundamentales: 1. Estudio de los entornos Open Data y la literatura al respecto; 2. Desarrollo de un proceso de integración y 3. Desarrollo de una Interface Gráfica y Analítica. Aunque en una primera fase se implementaron los procesos de integración mediante Java y Oracle y la Interface Gráfica con Java (jsp), en una fase posterior se realizó toda la implementación con lenguaje R y la interface gráfica mediante sus librerías, principalmente con Shiny. El resultado es una aplicación que provee de un conjunto de Datos Ambientales Integrados en Tiempo Real respecto a dos gobiernos muy diferentes en España, disponible para cualquier desarrollador que desee construir sus propias aplicaciones.