XI Jornadas de Usuarios de R

Madrid, 14, 15 y 16 de noviembre de 2019

http://r-es.org/XIjuR/

11jcomite-organizador@r-es.org

Libro de Resúmenes

Jueves, 14 de noviembre

16:30 - 17:30: Conferencia Max Kuhn

Aula/espacio: Auditorio Repsol; Moderador/Responsable: Carlos Ortega

Designing R Modeling Packages.

Max Kuhn (R Studio Inc.)

Resumen pendiente

Viernes, 15 de noviembre

09:00 - 09:55: Ponente invitado: Bernd Bischl

Aula/espacio: Salón de actos; Moderador/Responsable: Santiago Mota

MLR3.

Bernd Bischl (Universidad de Munich)

Resumen pendiente

10:00 - 11:00: Premio Joven I: Aplicaciones en Genética

Aula/espacio: Salón de actos; Moderador/Responsable: José Luis Cañadas

Informes Rmarkdown interactivos como resultado del análisis avanzado de datos ómicos.

Victoria Fornés Ferrer (Tau Analytics)

El manejo de datos ómicos requiere una minuciosa planificación debido a los voluminosos ficheros y los desmesurados recursos computacionales. R y Bioconductor ya cuentan con numerosas librerías que facilitan la importación y garantizan un correcto pre-procesado de los datos. Sin embargo, a pesar del avance de estas tecnologías y la ciencia, aún se siguen utilizando técnicas estadísticas clásicas que no garantizan la obtención de resultados óptimos y reproducibles. Esto provoca una aumento de la tasa de falsos negativos y una interpretación muy limitada de la información proporcionada. En esta comunicación oral se presentarán diversas técnicas estadísticas avanzadas de análisis de datos de metilación como la regresión beta o la regresión ordinal. Además, se hará uso de técnicas robustas como la regresión basada en rangos, especialmente recomendada en conjuntos de datos que presentan observaciones anómalas. Los resultados obtenidos serán contrastados a través de un análisis de sensibilidad en el que se aplicarán métodos de penalización como Elastic Net. Por último, pero no menos importante, esta comunicación también tiene por objetivo demostrar la utilización de la extensión Rmarkdown como alternativa a la presentación de resultados en formato .html. Ésta es una solución muy versátil que permite la representación de gráficos interactivos en 2D y 3D mediante la librería plotly, así como la incorporación de tablas dinámicas y exportables de los efectos estimados y la información del EPIC Manifest relativa a los resultados obtenidos.

Estimaciones del coeficiente de inbreeding genómico humano mediante la librería detectRUNS de R y otros programas bioinformáticos..

Luis Javier Sánchez Martínez (Universidad Complutense de Madrid), Candela L. Hernández (Universidad Complutense de Madrid); Abel Sánchez (Universidad Complutense de Madrid); Rosario Calderón (Universidad Complutense de Madrid).

Los matrimonios entre parientes biológicos constituyen un importante componente de la estructura marital de una población humana. El fenómeno consanguíneo todavía persiste en muchas sociedades, las cuales se caracterizan por unas reglas socioculturales arraigadas durante generaciones. El coeficiente de inbreeding (F) valora la probabilidad de homozigosis (autozigosis) en el genoma de un individuo. Tradicionalmente, las estimas del valor de F se han basado en el estudio de árboles familiares. Sin embargo, en estos últimos años se está produciendo un cambio de paradigma, el genotipado masivo de millones de SNPs (Single Nucleotide Polymorphism) a lo largo del genoma, constituye en la actualidad la tecnología molecular más novedosa y resolutiva. En el presente estudio, se ha estimado el coeficiente de inbreeding genómico en 142 individuos, autóctonos del sur de Iberia (Andalucía), sur de Portugal y de Marruecos (Bereberes). Para ello, se trataron los datos basados en 2.500.000 SNPs mediante distintos software bioinformáticos como PLINK y FSuite, además de la librería detectRUNS de R. Posteriormente, se calculó el coeficiente de correlación intraclase para cada una de las tres estimas del coeficiente de inbreeding genómico, tomando como referencia el estadístico calculado mediante la librería detectRUNS (denominado FROH). Las publicaciones recientes sobre este tópico, las cuales contienen diferentes modelos de simulación genómica, concluyen que el FROH es el estimador más poderoso para detectar la existencia de parentesco genético. Aplicando las tres metodologías, los valores promedio de los coeficientes de inbreeding genómico estimados fueron distintivos, destacando el obtenido por PLINK (FPLINK= 0.1229, FROH= 0.0535, FFSuite= 0.0162). El coeficiente de correlación intraclase presenta una mayor concordancia entre las estimas de FROH y FSuite que con respecto a PLINK [FROH vs. FFSuite =0.841 (IC95%: 0.784; 0.883) y FROH vs. FPLINK = 0.499 (IC95%: 0.363; 0.614)]. El presente estudio nos ha permitido contrastar la rentabilidad de las diferentes metodologías bioinformáticas disponibles hoy para analizar la consanguinidad humana desde una perspectiva genómica.

Computational inference for two-sex branching processes for X-linked recessive disorders.

Alicia León Naranjo (Universidad de Extremadura), Miguel González, Cristina Gutiérrez, Alicia León, Rodrigo Martínez

The evolution of the number of individuals carrying the alleles, R and r, of a gene linked to X chromosome has been described using a multitype two-sex branching process in (see [1]). The R allele is considered dominant and the r allele is supposed to be recessive and defective, responsible of a disorder. Hemophilia, red-green color blindness or the Duchenne and Becker's muscular dystrophies are examples of this kind of diseases. For this model we investigate the estimation of its main parameters from a Bayesian standpoint. Concretely, we apply the Approximate Bayesian Computation (ABC) methodology to approximate its posterior distributions. The accuracy of the procedure is illustrated and discussed by way of simulated examples developed with R.

Acknowledgements: The research has been supported by the Ministerio de Economía y Competitividad (grant MTM2015-70522-P), the Junta de Extremadura (grants IB16103 and GR18103) and the Fondo Europeo de Desarrollo Regional.

References: [1] González, M., Gutiérrez, C., Martínez, R.,~and Mota, M. (2016) Extinction proba-

bility of some recessive alleles of X-linked genes in the context of two-sex branching processes. In Branching Processes and their Applications, Lecture Notes in Statistics – Proceedings (del Puerto, I.M. et al., Eds), vol. 219, chapter 17. Springer-Verlag. DOI:~10.1007/978-3-319-31641-3

10:00 - 11:00: Análisis de datos

Aula/espacio: Ricardo Marín; Moderador/Responsable: Jorge Luis Ojeda

Oportunidades y limitaciones del análisis de texto con R: la discografía de Taylor Swift.

Ariane Aumaitre Balado (Instituto Universitario Europeo)

El análisis de texto automatizado ha abierto en los últimos años una ventana de oportunidades para el análisis de grandes cuerpos de texto que sería difícil estudiar de forma cualitativa. En R, el paquete tidytext ofrece herramientas para llevar a cabo tareas de procesamiento y limpieza de texto, análisis de sentimiento o relación entre distintos corpus. Pero más allá de las oportunidades de análisis que nos ofrecen estas herramientas, debemos ser conscientes de las decisiones que implica utilizarlas, así como de las limitaciones de las mismas.

Utilizando como estudio de caso la discografía de Taylor Swift, la presentación cubrirá el tipo de técnicas que ofrece tidytext y cómo llevarlas a cabo (siguiendo este esquema https://github.com/aaumaitre/taylor_swift), haciendo especial énfasis en el tipo de decisiones que se toman durante el análisis: ¿qué palabras frecuentes quedarnos, cuáles eliminar?¿qué es prescindible de nuestro texto y qué forma parte de su esencia?¿cómo elegir diccionarios?

Una parte de la presentación se dedicará a la evaluación crítica de los resultados y a las limitaciones del análisis de sentimiento automatizado. La utilización de canciones de Taylor Swift en lugar de textos más densos para este ejercicio nos permitirá detectar con mayor facilidad el tipo de problemas o sesgos que pueden surgir de esta técnica: es más fácil darnos cuenta de que una canción positiva que conocemos es categorizada como negativa por nuestro algoritmo que hacerlo con el texto de varias horas de debate parlamentario.

Análisi s d dato s.

Xavi de Blas (Asociación Chronojump)

Se analizará in situ un conjunto de datos sin conocer su procedencia. Las consecuencias que se deriven pueden ser muy severas.

Clasificación Automática de niveles de actividad en Acelerometría.

Jorge Luis Ojeda Cabrera (Universidad de Zaragoza), Jorge Luis Ojeda Cabrera, Jorge Marin-Puyalto, Jose Antonio Casajus

El uso de datos acelerométricos forma ya parte de la metodología fundamental a la hora de recabar información en áreas de investigación tan dispares como la deportiva, social, geológica o el área técnica. En particular, en el ámbito médico-deportivo, el uso de acelerómetros en los estudios relativos al desarrollo e intensidad de la actividad física ha aumentado notablemente en los últimos años, y más aún en gerontología.

El uso de este tipo de dispositivos para estudiar el desarrollo e intensidad de la actividad física en el área sanitaria resulta peculiar. Esta especificidad se debe, en primer lugar al objetivo que se busca:

determinar si el individuo hace mucho o poco ejercicio (p. ej. andar,correr, …). Pero también es distintiva la forma en que se mide: cuanto más activo está un individuo, cuanto más se mueve, más cambios en la aceleración se registran en un acelerómetro situado en la cintura o, preferentemente, en la muñeca. El que el nivel de actividad física, el objetivo de este tipo de estudios, se mida en base a la aceleración que esta actividad genera en partes del cuerpo, de una forma tan indirecta y el que esa relación pueda depender de del sujeto bajo análisis genera problemas adicionales, p.ej. cómo detectar periodos de inactividad, o en los que el paciente no lleva puesto el dispositivo (*Non Wear problem*).

En este trabajo, y con el ánimo de determinar los diferentes niveles de actividad por los que pasa un individuo cuando se le somete a un registro acelerométrico, se propone un modelo sencillo basado en variables ocultas para modelar la aceleración de la muñeca según los diferentes niveles de actividad por los que va pasando el sujeto. Los parámetros de dicho modelo se estiman mediante el uso del Algoritmo Expectation-Maximization (EM) que, como subproducto, permite clasificar diferentes periodos de tiempo según su nivel de actividad.

10:00 - 11:00: Gestión modelos y proyectos

Aula/espacio: Fdez. Huerta; Moderador/Responsable: Francisco Jesús Rodríguez Aragón

Project delivering with R packages.

Ernest Benedito (McKinsey & Company)

El objetivo de la presentación sería explicar una metodología para implementar proyectos de R como un paquete. Los puntos serían: • Cómo hacer el set up de un paquete de R de forma simple • Cómo estructurar de forma eficaz scripts, documentos de configuración y librerías necesarias • Cómo hacer la documentación

Gestión de la Capacidad en entornos Cloud mediante Análisis de Componentes Principales.

Jorge Pradas Moscardó (EDICOM), Ramón Ferrer Mestre

Dentro de un entorno cloud una de las gestiones más importantes se definir la capacidad de clientes que puede disponer cada una de las granjas (servidores). Tendremos que tener en cuenta muchas variables como el número de clientes, número de transacciones, errores registrados en el último período, etc… Con un análisis de componentes principales determinaremos que variables determinarán mejor nuestro sistema y dimensionaremos nuestros servidores de tal manera que no se exceda de los límites determinados. Usaremos la librería FactomineR para realizar el cálculo y Forecast para estimar la entrada de nuevos clientes y poder gestionar temporalmente los recursos. La interfaz que usaremos estará hecha con Shiny.

Gestión de modelos Python con Reticulate.

Francisco Jesús Rodríguez Aragón (Servicios Financieros Carrefour)

Actualmente los lenguajes claves para la modelización estadística y el desarrollo de aplicaciones empresariales son R y Python. Resulta por tanto habitual que en los equipos multidisciplinares que se encuentra habitualmente en el mundo profesional, aparezcan perfiles DS (Data Scientist) que tienen habilidad especial con uno o ambos de los anteriores lenguajes. Lo que se demuestra que resulta un error que redunda en ineficiencias es cuando se obliga a un DS especializado en R

que aprenda Python al mismo nivel que su lenguaje nativo y viceversa, aunque en ocasiones por restricciones técnicas los objetos serializados tienen que realizarse en uno u otro lenguaje. En este trabajo se presenta como la librería reticulate de R viene a solucionar el problema anteriormente citado, por una parte dicha librería permite interpretar código python creado externamente por un DS y devolver los correspondientes objetos dentro del entorno R (R-Studio), por otro lado también permite considerar objetos serializados .pkl de python y aplicarlos directamente a data frames generados dentro de un entorno R. Además al generar reticulate objetos de R, estos son perfectamente integrables con la conocidísima librería shiny permitiendo la construcción de aplicaciones web de modo muy rápido a la vez que permite a los DSs especializados en distintos entornos, la presentación y aplicación de sus resultados bajo un entorno único. En esta ponencia se muestra un ejemplo de un aplicativo shiny que ejecuta modelos serializados y construidos en entorno python que actualmente se están aplicando en la realidad para la identificación de los perfiles de gasto y comportamiento entre aquellos que acaban de contratar una tarjeta de crédito.

10:00 - 11:00: Aplicaciones interesantes de R

Aula/espacio: Florentino Sanz; Moderador/Responsable: Virgilio Gómez-Rubio

R en la empresa: cómo vender más disfraces integrando R.

Jesús Armand Calejero Román (Funiglobal)

Uno de los retos más importantes de los últimos años en Funidelia, e-commerce de disfraces y merchandising con presencia en en 32 países de tres continentes, ha sido potenciar su área de Data Science y desarrollar una plataforma de análisis y explotación de los datos integrando R como una nueva tecnología dentro del ecosistema ya existente. Para ello se debe dotar de agilidad al área para probar de forma independiente y rápida los experimentos que se llevan a cabo y desarrollar una factoría de datos robusta y eficiente. Para conseguirlo se han integrado distintas herramientas como RStudio Server, más orientadas a la ciencia de datos, o Gitlab o Jenkins, del ámbito de devops, en interacción continua con bases de datos tales como MySQL, Postgres o Hive y una infraestructura de almacenamiento de datos construida sobre Hadoop.

En esta presentación se muestra cómo se ha desarrollado el proyecto, desde el análisis inicial hasta su puesta en producción a través de distintas herramientas, es decir, cómo se escribe el código, cómo se gestiona en repositorios, el proceso de tests y su despliegue final.

simmer: Discrete-Event Simulation for R.

Iñaki Úcar (Universidad Carlos III de Madrid), Bart Smeets (dataroots)

The simmer package brings discrete-event simulation to R. It is designed as a generic yet powerful process-oriented framework. The architecture encloses a robust and fast simulation core written in C++ with automatic monitoring capabilities. It provides a rich and flexible R API that revolves around the concept of trajectory, a common path in the simulation model for entities of the same type.

11:30 - 12:55: Taller Max Kuhn

Aula/espacio: Salón de actos; Moderador/Responsable: Carlos Ortega

Designing R Modeling Packages.

Max Kuhn (R Studio)

R inherited informal conventions for modeling functions from the White Book (Statistical Models in S, 1991). However, many modeling packages, especially those developed by people new to R, don't follow these. This talk will discuss good and bad ways to create modeling packages based on a set of opinionated principles. An R package will be used to show how a high-quality modeling package can be made with little effort. Syntax from the tidymodels repositories will also be used for demonstration.

13:00 - 14:00: Premio Joven II: Metodología y otras aplicaciones

Aula/espacio: Salón de actos; Moderador/Responsable: Pedro Concejero

Interpoladores determinísticos espacio-temporales, series de tiempo y análisis de datos funcionales para el estudio y predicción de la precipitación en Cundinamarca y Bogotá D.C. durante el cuatrienio 2017-2020.

Diego Alejandro Malagón Márquez (Universidad Distrital Francisco Jose de Caldas Bogota Colombia), Carlos Eduardo Melo Martinez (Ingeniero Catastral y Geodesta, Ms. Economía, Doctor en Estadística-Universidad Distrital Francisco Jose de Caldas, Universidad Nacional de Colombia); Dilson David Ramirez Forero (Ingeniero Catastral y Geodesta-Universidad Distrital Francisco Jose de Caldas)

En el trabajo propuesto se abordó una metodología para el estudio y la predicción de la precipitación en el Departamento de Cundinamarca y la ciudad de Bogotá D.C, Colombia, a partir de registros de 133 estaciones meteorológicas del IDEAM (Instituto de Hidrología, Meteorología y Estudios Ambientales), comprendidas entre enero del 2010 y diciembre del 2016 con una frecuencia mensual, utilizando interpoladores determinísticos espaciotemporales, series de tiempo y análisis de datos funcionales, como una alternativa a los modelos numéricos y geoestadísticos más comunes, que, si bien son más robustos, presentan limitaciones en cuanto a la escala, complejidad e intervalos de predicción; utilizando para ello el software R, manipulando paquetes como geosptdb, fda, fda.usc, forecast y el modelado grafico geográfico con ggplot2 y spacetime. Se obtienen entonces predicciones espacio-tiempo para los años 2017, 2018, 2019 y 2020 (48 meses) a escala local y/o regional, con un buen nivel de detalle y de baja complejidad, incorporando además series de tiempo, como un complemento al proceso de interpolación, generando así un nuevo método competitivo respecto al uso únicamente de interpoladores determinísticos, pues, aunque genera un costo computacional mayor, se obtienen predicciones de precipitación con un menor error de predicción que pueden ser usadas como insumo fundamental en la planeación, el ordenamiento territorial y, la gestión y prevención del riesgo.

Using R for calculating efficiency in local governments.

Pedro José Martínez Córdoba (Universidad de Murcia), María-Dolores Guillamón (Universidad de Murcia); Bernardino Benito (Universidad de Murcia)

The document synthesizes the methodology for calculating efficiency and its determining factors in municipal public services. In addition, the waste collection service is analysed as an example of local public sector auditing and accounting research. The aim of this paper is to present the incorporation of R in this matter, detailing the different libraries used in R-studio. The multiple

advantages offered by R-studio for analysing, comparing and interpreting the results, making it possible to carry out the whole process in the same software, favour the research work.

rknn: agregando ordenaciones de vecinos en knn.

Noelia Rico (Universidad de Oviedo), Raúl Pérez-Fernández (Universidad de Oviedo); Irene Díaz (Universidad de Oviedo)

Rknn es un método de aprendizaje automático, destinado a resolver problemas de clasificación. Se trata de una extensión del conocido método knn pero en este caso la elección de los vecinos más cercanos se basa en el criterio de distintas distancias en vez de en una única distancia. Para ello, cada una de las funciones distancia elegidas por el usuario genera una ordenación de los vecinos más próximos. A continuación, estas ordenaciones son agregadas aplicando una regla de ordenación parametrizable. Esta implementación ha sido desarrollada para poder ser integrada con las funciones del paquete caret, lo cual facilita el entrenamiento, validación y comparación de los modelos entrenados con este método rknn.

13:00 - 14:00: Sociedad y cultura

Aula/espacio: Fdez. Huerta; Moderador/Responsable: Fran Ibáñez

La autocorrelación espacial está en todas partes.

Pelayo Arbués (Idealista/data), David Rey (CDO en Idealista/data)

El análisis estadístico de los residuos que cometen los modelos de aprendizaje nos permite extraer patrones sobre los que iterar en la mejora de dichos modelos. En este caso nos centramos en el estudio de la autocorrelación espacial de los modelos que utilizamos en la valoración de bienes inmuebles. Utilizando estadísticos como la I de Moran y los Test de Multiplicador de Lagrange podemos identificar aquellas especificaciones que se ajustan mejor al proceso generador de datos. Además de apoyarnos en el cálculo de dichos test estadísticos utilizamos la inspección visual de los datos y de los errores subyacentes utilizando H3, el mallado espacial basado en hexágonos desarrollado por Uber.

Usando R como el pilar de una infraestructura analítica en una organización cultural.

Luis Martinez-Uribe (Fundación Juan March), Fernando Martínez de Guzmán

La Fundación Juan March es una organización sin fines de lucro fundada en 1955 dedicada a la promoción de la cultura en España. Con sede en Madrid y dos museos en Palma y Cuenca, la fundación produce exposiciones, conciertos y series de conferencias. En Madrid, también alberga una biblioteca de investigación especializada en teatro y música contemporáneos españoles, ilusionismo y estudios curatoriales.

Es dentro de esta Biblioteca donde ha surgido una unidad de datos dedicada a proyectos de curación y análisis. Esta unidad es conocida como el DataLab y su objetivo es apoyar a la fundación con la difusión y preservación de sus activos digitales, así como a la toma de decisiones mediante el análisis de datos. Para lograr lo anterior, se ha establecido una infraestructura de datos en la nube con R como el pilar principal que desempeña un papel destacado en una variedad de tareas.

La infraestructura de datos comprende una capa de datos dedicada basada en MongoDB donde una colección orquestada de scripts de R captura y refina constantemente los datos de una variedad

de fuentes, como bases de datos internas, medios sociales, eventos de boletines, Google Analytics y otras bases de datos externas. Mediante varios paquetes de R, los datos se enriquecen al conectarse a servicios de inteligencia cognitiva y artificial que proporcionan procesamiento de lenguaje natural o capacidades de voz a texto.

El servidor RStudio proporciona un entorno de codificación común para los miembros del equipo de DataLab donde el acceso a la capa de datos y la reutilización de nuestro código es sencillo. Los cuadernos Jupyter con un kernel R también están configurados para probar, ejecutar y documentar análisis y visualizaciones particulares que ayudan a comunicar las ideas. Finalmente, se utilizan paneles de control con shiny para publicar KPI, métricas y predicciones, para ayudar a identificar tendencias, monitorizar el rendimiento y producir visualizaciones para video walls.

Big data, Fútbol y R. Casos Prácticos..

Jesús Lagos Milla (ORANGE)

En la charla abordaremos la aproximación que se está haciendo desde el mundo del fútbol al nuevo paradigma del dato, haciendo un recorrido por la captura de información por diferentes proveedores, procesado, entornos de análisis y ejemplos de uso mediante R. 1. Captura de Información, proveedores y casos de uso: a. Eventing b. Tracking c. Video d. Scouting e. Fuentes de información Open Source 2. El uso de R para el análisis de información: a. Librería soccergraphR que permite explotar los datos de Opta (proveedor de información de partidos de fútbol) b. Explotación de datos descargados desde la librería StatsbombR (Análisis FC Barcelona 2004-2012) c. Librería understatR que permite hacer scrapping de datos de partidos de fútbol de métricas avanzadas d. Librería FootballBadges para visualización de datos de understatR e. Ejemplos de visualización de datos con Shiny 3. Algunas líneas de trabajo: a. Detección de objetos. Scrapping Marca y análisis de aparición de escudos. [Google Colab] b. Clasificación de imágenes para el cálculo del xG c. Rotura de enlaces en el grafo. ¿cómo podemos romper tendencias de juego? Análisis postural. La orientación del cuerpo en el fútbol.

13:00 - 14:00: Encuestas y educación

Aula/espacio: Ricardo Marín; Moderador/Responsable: Guido Corradi

Psicometría con R aplicada a estudios sociológicos: análisis de encuestas de opinión pública.

Guido Corradi (Universidad Camilo José Cela), Airane Aumaitre (Instituto Universitario Europeo)

La psicometría es una rama de la psicología dedicada a la medición y modelización de constructos de carácter psicológico. Recientemente la sociología ha mostrado cierto interés en el uso de las herramientas que provee la psicometría. Para ello, R ofrece interesantes propuestas que permiten modelizar respuestas de encuestas en el marco de la teoría de rasgo latente, según la cual las respuestas a las preguntas de un cuestionario están causadas por variables no observables que se intentan modelizar. En esta charla se expondrán los básicos de diversas técnicas y paquetes de R que pueden ser útiles para el interesado en la modelización de respuestas a encuestas usando análisis factorial, teoría de respuesta al ítem y análisis de perfil latente a través de los paquetes psych, lavaan, ltm y mclust. Además, se proveerá de código y un ejemplo utilizando encuestas del CIS sobre actitudes hacia la familia y la igualdad de género.

Grafos interactivos con redes lineales y logísticas.

Modesto Escobar Mercado (Universidad de Salamanca), Luis Martínez-Uribe

Los grafos no solo han sido empleados para solucionar problemas topográficos y para representar estructuras sociales, sino también parar estudiar relaciones entre variables. Son bien conocidos los análisis de senderos (path analysis) y los modelos de ecuaciones estructurales. Ambos, sin embargo, estaban inicialmente restringidos al uso de variables cuantitativas. En esta presentación, se aborda cómo pueden también representarse las relaciones entre variables cualitativas, tal como ya lo hace el análisis de correspondencias, pero empleando los recursos técnicos del análisis de redes y de otras conocidas técnicas multivariantes como la regresión lineal y la logística. El análisis propuesto está basado en la realización de ecuaciones de regresión simultáneas y en la selección de aquellas relaciones con una relación estadística positiva y significativa. Con ello, se obtienen grafos en los que se destacan aquellas categorías que poseen una media neta en una serie de variables seleccionadas por encima del conjunto de la muestra. Para mejorar su potencial analítico a estos grafos se les dota de un potencial interactivo. La interacción gráfica comprende la selección de diversos atributos para el reconocimiento de los elementos analizados y la modificación de parámetros para centrarse en aquellas relaciones más fuertes o centrales. En la primera parte de la presentación, se abordarán los fundamentos matemáticos y estadísticos de estas representaciones, en la segunda parte, se describirá el paquete de R llamado netCoin, y se dará ejemplos de su uso interactivo en el análisis de encuestas y bases de datos.

Herramienta para el análisis de rendimiento académico del Grado en Ingeniería Informática en la UPV.

Cesar Ferri Ramírez (UPV), Artur De Osset Greño; Cèsar Ferri; Antonio Molina

Con el paso de los años se han matriculado cientos de estudiantes de toda índole, con variedad de rendimiento académico en el grado de ingeniería informática de la Escuela Técnica Superior de Ingeniería Informática de la Universitat Politècnica de València. De todos se guardan en bases de datos cierta información en el momento de su ingreso y toda su trayectoria dentro de la universidad. En este proyecto se ha desarrollado una herramienta web pensada para facilitar el análisis de toda esa información centrándose en el rendimiento, permitiendo focalizarse en las asignaturas o el alumnado, el filtrado de datos, incluso realizar ciertas predicciones. La herramienta se ha desarrollado en R. Gracias a la librería Shiny, es muy sencillo crear una aplicación web, permitiendo centrar los esfuerzos en el desarrollo de las módulos internos y características que permitan realizar un análisis de datos efectivo.

15:00 - 17:25: Talleres paralelos I

Aula/espacio: Ricardo Marín; Moderador/Responsable: Felipe Ortega

Aálisis de datos funcionales, regresión, clasificación y clustering funcional mediante la librería fda.usc.

Manuel Oviedo (Universidad de Santiago de Compostela)

El análisis de datos funcionales (AFD) es un campo de investigación muy activo durante los últimos años ya que aparecen de manera natural en la mayoría de los campos científicos: econometría (curvas de demanda o precio de la electricidad), medio ambiente (niveles de NOx), medicina (curvas de crecimiento), quimiometría (datos espectrométricos), etc.

Este taller ofrece una introducción al análisis de datos funcionales (FDA) en R utilizando el paquete fda.usc [1]. Este paquete lleva a cabo un análisis exploratorio y descriptivo de los datos funcionales, explorando sus características más importantes, como las mediciones de profundidad o la detección de valores atípicos funcionales, entre otros. También ayuda a explicar y modelar la relación entre una variable dependiente e independiente (modelos de regresión) y hacer predicciones. Se incluyen los métodos para la clasificación supervisada o no supervisada de un conjunto de datos funcionales con respecto a una característica de los datos. Puede realizar ANOVA funcional, pruebas de hipótesis, modelos de respuesta funcional y muchos otros.

Este curso se enfoca en: 1. Representación funcional de datos: • Definición de la clase fdata • Base fija y data-driven basis • Suavizado 2. Análisis exploratorio de datos funcionales. • Estadísticas de localización y dispersión. • Mediciones de profundidad y detección de valores atípicos. 3. Regresión funcional para la respuesta escalar. • Modelos funcionales lineales y no lineales. • Modelos funcionales generalizados. • Selección de variables 4. Clasificación funcional • Regresión binaria y multiclase • Basada en profundidades (DD-plot) • Aplicada a datos hiperespectrales • k-medias funcional

Install fda.usc package, https://cran.r-project.org/web/packages/ fda.usc R task view devoted to FDA: https://cran.r-project.org/web/views/FunctionalData.html

[1] Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: the R package fda. usc. J. Stat. Softw., 51(4):1–28. [2] Ramsay, J. and Silverman, B. (2005). Functional Data Analysis. Springer. [3] Ferraty, F. and Vieu, P. (2006). Nonparametric functional data analysis: theory and practice. Springer.

15:00 - 17:25: Talleres paralelos II

Aula/espacio: Fdez. Huerta; Moderador/Responsable: José Luis Cañadas

Herramientas de R para la investigación de mercados.

José Ignacio Casas Álvarez (Free-Lancer)

1. Por qué y para qué Buena parte del desarrollo y utilización de R en España se ha realizado en el ámbito académico, con una menor presencia en el campo de la investigación social y de mercados. El taller tiene como objetivo explorar este último terreno, mostrando el abanico de herramientas disponibles de una manera sucinta (dado el extenso campo a tratar) pero cubriendo áreas que suelen quedarse fuera de los manuales más habituales.
2. Contenido No se incluyen herramientas que, aunque utilizadas también en la investigación de mercados, están ya cubiertas de forma habitual en los manuales al uso. Por ejemplo: Análisis Exploratorio de Datos (EDA), modelos de regresión, Análisis de Componentes Principales (PCA), etc. Sin ánimo de exhaustividad, los contenidos del taller cubren las siguientes áreas: • Segmentación de mercados: uso de herramientas cluster • Channel Attribution & Sales Funnel (Canal de atribución & Embudo de ventas) • Basket Market Analysis & Association Rules (Análisis de la cesta de la compra & Reglas de asociación) • Costumer Value (Valor del cliente) & RFM (Recency, Frequency, Monetary value) • Conjoint Analysis • Técnicas de exploración de precios: Price Sensitivity Meter; Willingness To Pay, etc • Mapas perceptuales • Item Response Theory (Teoría de la respuesta al item)
3. Método de exposición Dado lo extenso de la materia tratada, la exposición se enfoca a un repaso esquemático de las herramientas disponibles resaltando cuándo y cómo utilizar cada

herramienta dependiendo del problema que se quiere solucionar. Si el tiempo lo permitiera (que no es previsible) se presentarían algunos ejemplos. El contenido de las transparencias, que estarán a disposición de los interesados, constituiría una especie de "cheat-sheet" sobre los temas tratados.

15:00 - 17:25: Talleres paralelos III

Aula/espacio: Salón de Actos; Moderador/Responsable: Emilio López Cano

Tips, trucos, y algunos paquetes para programacion eficiente en R.

Adolfo Alvarez (Collegium Da Vinci)

Aprender R es un proceso constante, no importa si estás dando tus primeros pasos o si eres una estrella del código, siempre hay espacio para mejorar. En este taller intentaremos descifrar qué es un código eficiente, cómo medir su eficiencia, cómo detectar cuellos de botella y cómo mejorar tu código. ¡Espero mostrar nuevas ideas, paquetes o trucos que quizas no conocías!

15:00 - 17:25: Talleres paralelos IV

Aula/espacio: Florentino Sanz; Moderador/Responsable: Leonardo Hansa

Taller de introducción al análisis reticular de coincidencias.

Luis Martinez Uribe (Universidad de Salamanca), Modesto Escobar(Universidad de Salamanca);Luis Martínez-Uribe (Universidad de Salamanca)

El principal objetivo del análisis reticular de coincidencias (ARC) es detectar qué sucesos, caracteres, objetos atributos o características tienden a aparecer conjuntamente en unos determinados escenarios. Su más remarcable característica es la combinación de múltiples análisis estadísticos multivariados con los análisis de redes basados en la teoría de grafos. Entre sus principales aplicaciones se encuentran el análisis de respuestas múltiples en cuestionario, el desarrollo de redes semánticas, el análisis de contenido, la minería de grandes bases de datos, el análisis de audiencias o el de cestas de compra. Todo ello se haría con dos paquetes: igraph, que es el clásico del análisis de redes, y con netCoin, que permite la generación de gráficos interactivos que proporcionan al análisis unas posibilidades exploratorias de las que carecen muchos otros análisis. Este curso pretende preparar a sus asistentes en los siguientes aspectos: a) conocimiento de los principales fundamentos de este análisis; b) utilización e interpretación de análisis a través de resultados en páginas web, y c) elaboración de los gráficos a partir de ficheros de bases de datos.

Temario 1.- Fundamentos del análisis reticular de coincidencias y de análisis de redes. Definiciones. Medidas y grados de coincidencias. Barras de coincidencias. Análisis de redes: nodos y adyacencias. Medidas de centralidad. Grafos: comunidades y disposiciones espaciales de los nodos. 2.- Interacción y uso de los grafos. Elementos de nodos y enlaces. Áreas de la herramienta de visualización: área reticular, área tabular, controles de la tabla de atributos, iconos del área tabular, controles del grafo, controles de fuerzas, controles de gráficos, nodos y enlaces. 3.- Construcción de los grafos. Elementos básicos de R. Importación de datos desde ficheros externos. La función dicotomizar. Funciones de gráficos: barras, barras condicionales, barras temporales y grafos. Control de las medidas. Elaboración de comunidades y distribuciones espaciales. El uso de imágenes.

17:30 - 18:00: Sesión de Posters y Café merienda

Aula/espacio: Anexo cafetería y columnata; Moderador/Responsable: Emilio López Cano

Cuando los árboles no te dejan ver el bosque: De la regresión logística al Random Forest.

Mercedes Ovejero Bruna (Sermes CRO / Universidad Complutense de Madrid), Sandra Toribio Caballero (Universidad Complutense de Madrid)

Uno de los análisis de datos más empleados en Psicología es la regresión logística. Este análisis requiere que no haya colinealidad entre las variables predictoras. Los algoritmos Random Forest permiten trabajar de manera flexible en casos de colinealidad. El presente trabajo supone la aplicación práctica de un modelo de regresión logística y el estudio de su rendimiento clasificador comparándolo con diferentes paquetes de R que permiten implementar algoritmos Random Forest: randomForest, randomForestSRC, ranger y Rborist. La base de datos pertenece a una tesis doctoral cuyo objetivo es clasificar como caso/control a una muestra de mujeres, conociendo sus puntuaciones en varias pruebas. La base de datos consta de 368 registros con 38 variables, siendo una de ellas la variable de clasificación y el resto predictoras. Los resultados muestran que: (1) El modelo de regresión logística tiene una peor capacidad clasificatoria debido a la colinealidad entre las variables predictoras (AUC=0.736). (2) Los modelos Random Forest consiguen mejorar la capacidad clasificatoria en el set de validación con independencia de los paquetes empleados (AUC$\approx$0.950). (3) Excepto en el caso del paquete Rborist, el porcentaje de errores de clasificación es similar. (4) El tiempo de ejecución de las funciones implementadas en ranger y randomForestSRC es inferior comparado con randomForest y Rborist. En conclusión, los algoritmos Random Forest se presentan como una solución alternativa en los casos en los que los supuestos de la regresión logística no se cumplan, especialmente en el caso de la colinealidad, problemática que afecta con mucha frecuencia en Psicología. No obstante, es imprescindible que los datos cumplan una serie de requisitos, por ejemplo, la distribución de la variable dependiente y la fiabilidad o precisión de los instrumentos utilizados. Además, es muy importante conocer los hiperparámetros de las funciones de Random Forest para evitar errores de ajuste.

RJSplot y D3GB: Gráficos interactivos para la visualización de datos biológicos en R..

David Barrios (Servicio de Bioinformática. Nucleus. Universidad de Salamanca.), Carlos Prieto (Servicio de Bioinformática. Nucleus. Universidad de Salamanca.)

Los avances en métodos de visualización de datos están generando nuevos métodos para la generación de gráficos interactivos. Estos gráficos permiten mejorar la exploración e intepretación de datos biológicos pero su generación requiere de un conocimiento avanzado de librerías gráficas. Para facilitar la visualización dinámica de datos en R, se están desarrollado paquetes que permiten la generación de gráficos interactivos mediante la llamada a funciones. Pese a su potencial y utilidad, R cuenta con un número limitado de paquetes que permite la generación de gráficos interactivos dirigidos a su aplicación en bioinformática y genómica. El proyecto que hemos desarrollado, une el poder de análisis de R con las características avanzadas de generación de gráficos interactivos de JavaScript en dos paquetes: RJSplot y D3GB. Estos paquetes permiten la generación sencilla de gráficos interactivos en R, proporcioanando nuevas capacidades de visualización, y contribuyendo al avance de los métodos de análisis bioinformático. En la actualidad, hemos desarrollado 17 gráficos interactivos y un navegador genómico que están disponibles en los paquetes RJSplot y D3GB. Estos gráficos son multiplataforma y se pueden ver mediante un navegador Web. Los paquetes están disponibles en sus páginas web (http://rjsplot.net y http://d3gb.usal.es) y en CRAN.

Aplicaciones Shiny para la enseñanza de la Estadística en Enseñanzas Medias.

Teresa González Arteaga (Universidad de Valladolid), Elia Muñoz Alonso

La Estadística posee un papel primordial en el desarrollo de la sociedad actual. Por ello, la enseñanza de la Estadística ha venido adquiriendo un mayor relieve en los currículos de Matemáticas en las Enseñanzas Medias del que se presentaba tradicionalmente. Utilizar una herramienta informática como apoyo para el estudio de esta parte de las materias de Matemáticas contribuye a mejorar el aprendizaje, motivar a los alumnos, aumentar su interés e incitarlos a que se sientan atraídos por esta ciencia. En este trabajo se muestran tres aplicaciones web interactivas, desarrolladas con Shiny, que recogen los contenidos de Estadística incluidos en las materias de Matemáticas en las Enseñanzas Medias.

R y Shiny al rescate de los tasadores: implementación de un algoritmo experto para modelos de valoración automática.

Emilio López Cano (Universidad de Castilla-La Mancha), Beatriz Larraz-Iribas (Sección MECYDES Instituto de Desarrollo Regional de la UCLM); José Luis Alfaro-Navarro (Facultad de Ciencias Económicas y Empresariales de la UCLM); Esteban Alfaro-Cortés (Sección MECYDES Instituto de Desarrollo Regional de la UCLM); Noelia García Rubio (Facultad de Ciencias Económicas y Empresariales de la UCLM); Matías Gámez Martínez (Sección MECYDES Instituto de Desarrollo Regional de la UCLM)

Los Modelos de Valoración Automatizada (AVM) permiten la valoración masiva de bienes inmuebles de forma rápida y sencilla. Esta valoración es necesaria en el sector bancario, no sólo a efectos contables, sino también para el cumplimiento de la normativa. Tradicionalmente, los tasadores con conocimientos especializados hacen esas valoraciones siguiendo algunas prácticas y normas comunes. Este método experto conduce a un conjunto de los llamados "comparables" (o más coloquialmente, testigos), es decir, propiedades cercanas con características similares, para valorar una propiedad. El surgimiento de Big Data y otros avances de las tecnologías de la Ciencia de Datos ha desencadenado un gran número de nuevos enfoques que utilizan herramientas estadísticas, como la regresión, los métodos basados en árboles o las redes neuronales, entre otros. Sin embargo, un enfoque de algoritmo experto resulta útil y más comprensible para las partes interesadas del mercado inmobiliario. En este trabajo se presenta la implementación de un algoritmo experto en R que imita el comportamiento de los tasadores humanos que buscan propiedades comparables en una base de datos depurada. El conjunto de normas se ha establecido después de un análisis exhaustivo en cooperación con una importante empresa de tasación. Estas reglas incluyen la detección de valores atípicos, la clasificación de datos de calidad y la estimación basada en la distancia. Los scripts están integrados en aplicaciones Shiny que permiten a los analistas evaluar la precisión del método y a los tasadores ejecutar el AVM sin problemas en una propiedad determinada, comprobando todos los pasos del método de estimación.

Bip4Cast: some advances in mood disorders analysis with R.

Victoria Lopez Lopez (Universidad Complutense de Madrid), Pavel LLamoca (Universidad Complutense); Diego Urgelés (Hospital Ntra Sra de la Paz); Milena Cukic (3EGA)

Hoy en día, hay proyectos en el área de la salud mental que están respaldados por dispositivos tecnológicos que mejoran la eficiencia de los tratamientos al permitir sin esfuerzo la recolección de indicadores biológicos y psicológicos de los pacientes. Nuestro proyecto se centra en descubrir

la relación entre los indicadores proporcionados por los datos recopilados. El objetivo principal es encontrar esos patrones comunes que podrían desencadenar una crisis de manía o depresión. La mayoría de los datos provienen de dispositivos portátiles, pero también los datos recopilados por el psiquiatra en consulta médica son parte del análisis. En la consulta médica se recoge el indicador de crisis, que es el indicador básico de nuestro proyecto. Además, para enriquecer el análisis, se recopilan otros datos sobre el estado de ánimo del paciente (consultoría médica y dispositivos portátiles). Este proyecto se está desarrollando junto con el hospital Ntra. Sra. de la Paz (España) donde algunos pacientes ya forman parte de la investigación experimental. Hay algunos conjuntos de datos recopilados en etapas anteriores y datos recopilados por dispositivos médicos en las pruebas (por ejemplo, Actigraph, ver GeneActiv, 2019). Dado que todos estos datos provienen de diferentes fuentes, hay una etapa difícil trabajando en la integración de datos. El objetivo es construir una estructura simple que contenga los indicadores de crisis y que facilite el análisis y la aplicación de un análisis de Machine Learning. Se planea desarrollar esos objetivos utilizando R.

Estudio de la eficiencia de los aeropuertos españoles a través del método DEA mediante el programa R.

Úrsula Faura-Martínez (Universidad de Murcia), Javier Cifuentes-Faura (Universidad de Murcia)

En el presente trabajo se realiza un estudio sobre la eficiencia de los aeropuertos españoles para el año 2018. Empleando el paquete estadístico R se calcula, en una primera etapa, la eficiencia aplicando el método DEA. Como inputs se han seleccionado variables que representan aspectos físicos de los aeropuertos como el número de puertas de embarque o el área de la pista de aterrizaje y despegue, entre otros. Entre los outputs, se encuentran el número de pasajeros, las toneladas de carga y el número de operaciones efectuadas en total por cada aeropuerto de la red aeroportuaria de AENA. Se trata, por tanto, más de conocer la eficiencia física que la financiera. A partir de esta técnica se puede conocer también en qué aspectos concretos debe mejorar cada aeropuerto para que le permita ser lo más eficiente posible y determinar en qué cuantía debería hacerlo. En una segunda etapa se realiza una regresión truncada para determinar qué variables influyen en la eficiencia de los aeropuertos. Previamente se aplica un bootstrapping con 2000 iteraciones para corregir el posible sesgo de los resultados del DEA para la regresión de la segunda etapa.

Web didáctica de análisis multivariante con R.

Ursula Faura Martínez (Universidad de Murcia), Fuensanta Arnaldos García (Universidad de Murcia); Mª Teresa Díaz Delfa (Universidad de Murcia); Lourdes Molera Peris (Universidad de Murcia); Isabel Parra Frutos (Universidad de Murcia); Juan José Pérez Castejón (Universidad de Murcia)

Dada la utilidad de los métodos multivariantes en el análisis de datos socioeconómicos, y con objeto de favorecer su aplicación por parte de los estudiantes de titulaciones relacionadas con la economía y la empresa, un grupo de profesores de la Universidad de Murcia nos hemos planteado la construcción de una web didáctica que sirva de guía en este contexto. Los materiales se han diseñado pensando en un aprendizaje autónomo, utilizando datos reales y ejemplos de interés para los estudiantes de estas titulaciones. Con esta idea en mente uno de los principales objetivos en el diseño ha sido emplear un formato visual, atractivo e interactivo, que permita al estudiante evaluar sus avances. Se ha utilizado R en todo el proceso de construcción de la web didáctica: la estructura del sitio web se gestiona con R Markdown, la aplicación de las técnicas se lleva a cabo con paquetes y funciones de R apropiados (importando los datos de la fuente correspondiente, en caso necesario) y la autoevaluación de los conocimientos adquiridos se resuelve con aplicaciones interactivas Shiny.

De momento, se ha elaborado material para las cuatro técnicas exploratorias de interdependencia más usuales (análisis de conglomerados, de componentes principales, de correspondencias simple y de correspondencias múltiple). Para cada una de ellas se ha creado un sitio web con un formato común que incluye cinco páginas: (1) una ficha, con los objetivos de la técnica, los requisitos de aplicación y posibles usos, junto con las funciones y paquetes de R necesarios; (2) un resumen teórico con los aspectos más relevantes de la técnica; (3) un caso práctico resuelto, que proporciona una orientación sobre el uso de la técnica y la interpretación de los resultados, e incluye un enlace para descargar los datos usados; (4) un caso práctico propuesto, para reforzar el aprendizaje del estudiante; y (5) una autoevaluación con cuestiones referidas al caso práctico propuesto, generadas aleatoriamente a partir de una batería de preguntas, lo que permite múltiples evaluaciones y que se corrigen en línea.

Visualizando secuencias de resolución en actividades de estadística aplicada.

Vanessa Serrano Molinero (IQS - Univ. Ramon Llull), Jordi Cuadros (IQS - Univ. Ramon Llull); Víctor León (IQS - Univ. Ramon Llull)

La resolución de problemas es un elemento fundamental en el aprendizaje de la estadística. Especialmente cuando estos problemas son abiertos, la evaluación del trabajo realizado por los estudiantes no es una tarea sencilla. Facilitar esta evaluación es lo que pretende el modelo que estamos desarrollando. El modelo propuesto comienza cuando se les pide a los alumnos que resuelvan una actividad semi-abierta utilizando una versión modificada de R Commander que traza sus acciones. Posteriormente, estas acciones se pueden visualizar y analizar utilizando un panel de control Shiny. A través de esta interfaz, se puede evaluar el trabajo de los estudiantes comparándolo con hitos de observación previamente establecidos, es decir, logros importantes o posibles errores en la resolución de la actividad. Hasta el momento se identificaba principalmente si se habían alcanzado estos elementos de observación en cualquier momento durante el desarrollo de la actividad. Recientemente, sin embargo, se han agregado nuevas visualizaciones para analizar la secuencia y la repetición de los elementos de observación, aproximándonos así de una forma más clara a lo que los estudiantes pueden estar entendiendo y dónde se pueden encontrar las principales dificultades. En este póster se presentan los resultados de la aplicación de este enfoque para analizar una actividad que se ha llevado a cabo en un primer curso de estadística aplicada a nivel universitario.

RAAPS: A Shiny App for Risk Assessment around Pollution Sources.

Virgilio Gómez Rubio (Universidad de Castilla-La Mancha), J. L. Gutiérrez Espinosa (Departamento de Matemáticas, E.T.S. de Ingenieros Industriales - Albacete, Universidad de Castilla-La Mancha, Spain), F. Palmí-Perales (Departamento de Matemáticas, E.T.S. de Ingenieros Industriales - Albacete, Universidad de Castilla-La Mancha, Spain), R. Ramis-Prieto (Environmental and Cancer Epidemiology Unit, Carlos III Institute of Health, Madrid, Spain and Consortium for Biomedical Research in Epidemiology & Public Health, CIBER Epidemiología y Salud Pública - CIBERESP, Spain), J. M. Sanz-Anquela (Cancer Registry and Pathology Department, Hospital Universitario Príncipe de Asturias, Madrid, Spain and Department of Medicine and Medical Specialties, Faculty of Medicine, University of Alcalá de Henares, Madrid, Spain) and P. Fernández-Navarro (Environmental and Cancer Epidemiology Unit, Carlos III Institute of Health, Madrid, Spain and Consortium for Biomedical Research in Epidemiology & Public Health, CIBER Epidemiología y Salud Pública - CIBERESP, Spain)

In this work we propose a method to identify several possible sources of exposure to pollutants that

might be producing negative health effects, as this is a common concern in Public Health. Assessing exposure to polluting industries often requires specific data and statistical methods to be addressed properly. For this reason, a protocol is required that takes advantage of regularly collected data, such as cancer registers, official cartography and registers of polluting industries. Pollution source investigations are prone to suffer from the Texas sharpshooter fallacy, i.e., assessing risk about a particular pollution source when an increased risk about it has already been observed. Here, we show how to conduct a spatial analysis of Public Health data to assess exposure to pollution sources using the R programming language and we propose new lines for future work and the development of a Shiny app.

18:00 - 18:30: Sesión relámpago I

Aula/espacio: Ricardo Marín; Moderador/Responsable: Virgilio Gómez-Rubio

Indicadores de riesgo en el rendimiento académico en grados de ingeniería..

Carlos de la Calle Arroyo (UCLM), Licesio J. Rodríguez-Aragón

Durante cuatro cursos académicos se han venido recopilando indicadores sobre el rendimiento académico de alumnos matriculados en primero de ingeniería en la Universidad de Castilla-La Mancha (Campus de Toledo). Parte de estos indicadores son proporcionados por los estudiantes de forma voluntaria a través de encuestas y cuestionarios. También se recopilan varios indicadores de calificaciones de las diferentes actividades evaluables desarrolladas a lo largo del curso.

Este trabajo persigue proporcionar un indicador de riesgo que nos permita predecir la probabilidad de que el alumno abandone o no consiga superar la asignatura. En base a la información disponible se persigue poder alertar a lo largo del curso sobre estrategias de estudio o comportamientos que, según la experiencia pasada, comprometen los resultados académicos de los alumnos.

Los análisis y estudios realizado, aún en desarrollo, se han centrado en el uso de técnicas estadísticas a menudo relacionadas con el machine learning, como son el PCA, K-means y otras técnicas de clustering, árboles de decisión y random forests, regresión, SVM, etc. En general, los análisis se han realizado de mayor a menor complejidad, de manera que sirvieran tanto de herramienta de clasificación/predicción como de método para comprender las relaciones entre variables. Se ha trabajado también en representaciones gráficas de los análisis, y una vez concluido el estudio se producirá un reporte y aplicativo en markdown y shiny respectivamente, incluyendo las conclusiones y facilitando la exploración de los y procesos por parte de otros.

R para la creación de un repositorio de vídeos sobre matemáticas, estadística y econometría de Universidades Españolas.

Jaime Pinilla Domínguez (Universidad de Las Palmas de Gran Canaria), Christian González-Martel (ULPGC); Miguel Ángel Negrín (ULPGC); José María Pérez-Sánchez (ULPGC)

Los materiales audiovisuales constituyen una herramienta educativa de gran utilidad. Evidencia de ello es el enorme crecimiento de la oferta de vídeos formativos en diferentes plataformas web, pero sin un criterio mínimo de calidad y con gran variedad de títulos y etiquetas. Encontrar un vídeo de calidad a través del buscador de YouTube requiere de cierta fortuna que no siempre da el resultado deseado. La mayoría de las universidades españolas cuentan con un canal institucional en YouTube, sin embargo la ausencia de un repositorio único hace que el material se disperse y sea sólo utilizado por los alumnos de cada universidad En este trabajo se desarrolla una herramienta web que sirve de

recopilación de vídeos educativos para las asignaturas de matemáticas, estadística y econometría de las titulaciones de Economía y Empresa de las 84 universidades españolas. Para ello entre otros paquetes utilizamos Blogdown, un paquete de R desarrollado por el creador de knitr y bookdown. El paquete se ejecuta utilizando un generador de sitio estático llamado "Hugo" que luego alojamos en Netlify. Se realizó una búsqueda exhaustiva de vídeos en la plataforma YouTube que llevasen las palabras 'matemáticas', 'estadística', 'econometría', 'ADE' o 'Economía' más el nombre o acrónimo de la universidad utilizando operadores de búsqueda booleanos. Se excluyen aquellos vídeos que no tengan referencias a la universidad como que no esté alojado en el canal oficial, no aparezca el nombre o logotipo de la universidad en el nombre del canal o descripción. La gestión de las etiquetas de cada vídeo se llevó a cabo mediante algoritmos de clustering implementados también en R. Nuestro repositorio web permite además de acceder al listado de enlaces y visualización del video, una búsqueda interactiva dentro del banco de recursos en función de diferentes criterios como pueden ser universidad, materia, tópicos tratados, etc.

R-Bingo.

Irene Hernández Martínez (Universidad de Murcia), Francisco Javier Ibáñez López (Universidad de Murcia)

Con motivo de la XVIII edición de la Semana de la Ciencia y la Tecnología de la Región de Murcia, que se celebrará en Murcia del 8 al 10 de noviembre de 2019, y que este año conmemora el año Internacional de la Tabla Periódica de los Elementos Químicos, desde la Sección de Apoyo Estadístico del Área Científica y Técnica de Investigación de la Universidad de Murcia hemos querido participar elaborando una sencilla aplicación shiny en la que los escolares y el público en general presente podrá descubrir los elementos de la tabla periódica mediante el juego, a través de un bingo elaborado con los elementos en vez de números. En la aplicación se ha contado con los elementos químicos principales de la tabla (90 elementos), presentados en su posición original y con los colores a imagen y semejanza de la tabla periódica expuesta en la fachada de la Facultad de Química de la Universidad de Murcia, descrita por su decano, el profesor D. Pedro Lozano, como "la más grande del mundo". Con una sencilla interfaz, aleatoriamente se van seleccionando elementos de uno en uno que se tiñen de rojo y además, se pueden consultar los últimos cinco elementos elegidos. Al mismo tiempo, se han elaborado también cartones con la ayuda del paquete "kableExtra", que muestran un total de 15 elementos siguiendo la disposición original de los cartones de bingo, mediante una división y clasificación de los todos los elementos de la tabla en 9 grupos con 10 elementos cada uno. En todo el proceso, se ha utilizado R a través de su GUI RStudio mediante la creación de elaboradas funciones buscando la adecuación de los elementos, la tabla periódica y los cartones a las reglas del juego.

Proyecto de adaptación de R/exams a la plataforma Sakai.

Fuensanta Arnaldos García (Universidad de Murcia), Jesús María Méndez Pérez (Universidad de Murcia); José Antonio Palazón Ferrando (Universidad de Murcia); Valentina Alacid Cárceles (Universidad de Murcia); Fuensanta Arnaldos García (Universidad de Murcia); María Victoria Caballero (Universidad de Murcia); Mª Teresa Díaz Delfa (Universidad de Murcia); Úrsula Faura Martínez (Universidad de Murcia); Lourdes Molera Peris (Universidad de Murcia); Isabel Parra Frutos (Universidad de Murcia); Nicolás Ubero Pascal (Universidad de Murcia)

El paquete `exams` de R permite la generación de distintos tipos de materiales de evaluación a partir de ficheros de texto escritos en `Latex` o lenguaje de marcas (`markdown`). Sus principales ventajas

son las siguientes: (1) Generación de PDFs para exámenes escritos clásicos, que incluyen un sistema para escaneado de las respuestas y evaluación automática; (2) elaboración y personalización de ficheros de distinto tipo (PDF, HTML, Docx, ODF…), con plantillas flexibles para los documentos; y (3) importación para muchos de los distintos Learning Management System (LMS) disponibles en el mercado (Moodle, Blackboard, OLAT, Ilias…), incluyendo ARSnova (Audience Response System) que consiste en un sistema de votación directa para el "aula invertida". El objetivo del proyecto es incorporar al paquete `exams` la funcionalidad necesaria para generar ficheros que puedan ser importados en Sakai. Básicamente, a partir de la función `exams2qti12` existente en el paquete se está trabajando en la construcción de una nueva función, `exams2sakai`, que incluya los cambios apropiados. Aunque el proyecto se encuentra todavía en las primeras fases de desarrollo, con las modificaciones realizadas actualmente ya es posible importar en Sakai preguntas del tipo Single Choice (schoice). Además, también nos planteamos desarrollar algunas plantillas que puedan ser de utilidad para distintas materias. Se trata de un proyecto de bastante utilidad, puesto que son muchas las universidades cuyo campus virtual se basa en el software educativo Sakai, en particular la Universidad de Murcia. El proyecto está abierto a cualquier persona que quiera colaborar.

18:00 - 18:30: Sesión relámpago II

Aula/espacio: Fdez. Huerta; Moderador/Responsable: Antonio Maurandi

AeRobiology: análisis de datos biológicos en el aire con R.

Jesús Rojo Úbeda (Universidad de Castilla-La Mancha), Antonio Picornell (Universidad de Málaga, España); Jose Oteros (Zentrum Allergie und Umwelt, Alemania)

La aerobiología estudia el contenido en el aire de organismos o partículas de origen biológico como son granos de polen, esporas de hongos o bacterias, entre otros. Esta labor presenta un gran interés desde diferentes puntos de vista como la salud pública, la agronomía o el medio ambiente. La gran cantidad de datos registrados durante décadas y el creciente avance de muestreadores automáticos requieren el uso de nuevas herramientas de análisis de mayor eficiencia que permitan la automatización de tareas. El paquete AeRobiology implementado en R se ofrece como solución para la interpretación de datos con claro comportamiento estacional como son concentraciones de polen y de esporas en la atmósfera. El paquete contiene numerosas funciones que se estructuran en tres importantes secciones del análisis de datos aerobiológicos como es la verificación de los datos (filtrado de datos, control de calidad de los datos, interpolación de datos ausentes, etc.), el cálculo de los principales índices utilizados en este ámbito científico (índice de la estación polínica, calendarios de exposición, tendencias de series temporales, entre otros) y la visualización de los resultados (numerosas opciones de visualizaciones en gráficos estáticos e interactivos). Más información sobre el paquete se puede consultar en Rojo, Picornel & Oteros (2019) Methods in Ecology and Evolution, https://doi.org/10.1111/2041-210X.13203. Los autores de este paquete de R, desde su incorporación al repositorio CRAN, han llevado a cabo diversas tareas de divulgación del paquete para fomentar su utilización en el ámbito científico en el cual se ha desarrollado. En este sentido se han impartido cursos presenciales y online relacionados con esta disciplina del conocimiento, el paquete se ha presentado en congresos de ámbito internacional, y se ha proporcionado el tutorial completo del paquete online en la siguiente dirección web: http://rpubs.com/Picornell/AeRobiology

Reducción dimensional clustering y visualización interactiva con looking4clusters: aplicación al análisis de expresión de célula única..

Angela Villaverde-Ramiro (Servicio de Bioinformática, Nucleus, Universidad de Salamanca.), David Barrios (Servicio de Bioinformática, Nucleus, Universidad de Salamanca); Carlos Prieto (Servicio de Bioinformática, Nucleus, Universidad de Salamanca).

Looking4clusters es un nuevo paquete de R que realiza una visualización interactiva de los datos para facilitar la determinación de clusters y la clasificación de muestras. Los datos de entrada se proyectan en representaciones bidimensiones mediante la aplicación de métodos de reducción de dimensionalidad como: PCA, MDS, t-SNE, UMAP, NMF. La representación de los datos en una misma interfaz con gráficos interconectados, permite identificar desde diferentes perspectivas los posibles agrupamientos. El paquete integra también la aplicación de técnicas de clustering no supervisado cuyos resultados se pueden visualizar interactivamente en la interfaz gráfica. Como resultado, el paquete genera una página Web interactiva desarrollada con JavaScript que se puede explorar fácilmente con cualquier navegador Web. Además de la visualización, esta interfaz permite seleccionar manualmente los grupos, identificar los clusters encontrados en las técnicas de clustering no supervisado, generar nuevos gráficos que muestren valores de las variables de entrada, localizar muestras y comparar visualmente los agrupamientos encontrados. Como ejemplo de uso, hemos analizado datos de expresión de experimentos de célula única en los que la identificación de grupos y la clasificación de muestras son procesos clave para la obtención de buenos resultados. La aplicación de looking4clusters en datos de célula única, permitió determinar los tipos celulares presentes en el experimento y realizar una clasificación de cada célula secuenciada a un grupo celular.

Control de la Producción en queserías del grupo Lactalis.

Marta Alonso Abalde (Grupo Lactalis), Inés; Carlos; Pablo; Miguel Angel (todos pertenecientes al grupo Lactalis)

Realizaremos una presentación breve sobre las herramientas de R que utilizamos para el control de procesos, seguimiento de la variabilidad, análisis de series u optimización de procesos como el corte de quesos. También mostraremos la integración de R con Power Bi de la cual nos servimos para el seguimiento diario/semanal y mensual de los parámetros técnicos de control al diseñar una interfaz que usuarios ajenos a R sean capaces de utilizar sin tener que tener un conocimiento previo. Finalizaremos la presentación mostrando brevemente un estudio realizado para el seguimiento y optimización de un proceso concreto.

FSinR: Un paquete extenso para la selección de características.

Francisco Aragón Royón (Universidad de Granada), Alfonso Jiménez Vílchez (Universidad de Córdoba); Antonio Araúzo Azofra (Universidad de Córdoba); José Manuel Benítez Sánchez (Universidad de Granada)

La selección de características es el proceso que consiste en seleccionar una serie de variables relevantes para la construcción de modelos. Este proceso es actualmente una parte clave del Aprendizaje Automático y su aplicación presenta un gran impacto en el rendimiento de los modelos, debido a que se prescinde de las variables sin carga predictiva o redundantes que pueden añadir ruido y empeorar el ajuste de los modelos, además de incrementar considerablemente el tiempo de cómputo y la complejidad. La principal dificultad que presenta la selección de características es que es un proceso de gran complejidad, es un problema NP-completo, y no existe una solución universalmente válida para todos los problemas. Por esta razón siguen siendo interesantes las nuevas propuestas y aportaciones. En el repositorio CRAN existen paquetes que abordan el problema de la selección de características, pero se centran en métodos concretos y no presentan una revisión amplia de los

métodos existentes. Para cubrir las carencias de los paquetes actuales hemos desarrollado el paquete FSinR. El paquete contiene una gran variedad de los métodos de filtro más ampliamente usados en la literatura (métodos como Gain ratio, Gini Index, Relief, información mutua, incertidumbre simétrica, medidas de consistencia, RFSM, etc.), así como una extensa variedad de métodos de wrapper. Para generar los métodos de wrapper nos valemos del paquete Caret, con sus casi 300 modelos disponibles de clasificación y regresión. Además, el paquete implementa también diversas estrategias de búsqueda (búsqueda en profundidad, anchura, secuencial, tabú, algoritmos genéticos, de colonias de hormigas, Las vegas, etc.) que se combinan con los métodos anteriores para guiar el proceso de selección de características. Por lo tanto, el paquete se destaca por ser primero en ofrecer en un mismo paquete una herramienta extensa, completa y fácil de usar para realizar de selección de características.

18:00 - 18:30: Sesión relámpago III

Aula/espacio: Florentino Sanz; Moderador/Responsable: José Luis Cañadas

Quantity Calculus for R.

Iñaki Úcar (Universidad Carlos III de Madrid), Edzer Pebesma (Universität Münster)

A quantity is a measurable property of a phenomenon, body, or substance, that is composed of (1) a value (a number representing the measurand), (2) a measurement unit (or magnitude), and (3) a measurement error (or uncertainty). Traditionally, computational systems have treated these three components separately. The R Quantities project, funded by the R Consortium, integrates the R packages 'units', 'errors' and 'quantities' into a complete quantity calculus system, which provides measurement units and uncertainty in R vectors, matrices and arrays, with automatic conversion, derivation, simplification, propagation and reporting.

Descarga de datos del Instituto Nacional de Estadística con R usando el servicio API JSON.

Daniel Redondo Sánchez (Instituto de Investigación Biosanitaria ibs.GRANADA, Escuela Andaluza de Salud Pública, CIBERESP), Miguel Ángel Luque Fernández (Instituto de Investigación Biosanitaria ibs.GRANADA, Escuela Andaluza de Salud Pública, CIBERESP), Miguel Rodríguez Barranco (Instituto de Investigación Biosanitaria ibs.GRANADA, Escuela Andaluza de Salud Pública, CIBERESP), Pablo Fernández-Navarro (Centro Nacional de Epidemiología, CIBERESP), María José Sánchez Pérez (Instituto de Investigación Biosanitaria ibs.GRANADA, Escuela Andaluza de Salud Pública, CIBERESP)

En este trabajo describimos un método de descarga de información del Instituto Nacional de Estadística (INE) usando R. El código es totalmente reproducible y está disponible en un repositorio de GitHub: https://github.com/danielredondo/INE_R

Utilizamos el servicio API (Application Programming Interface) del INE para realizar la tarea de conexión e intercambio de datos. En primer lugar, obtenemos la URL (dirección web) válida para la descarga, en función del tipo de información a descargar (por ejemplo, si es una tabla con número definido, o si es un fichero PCAxis). Después, procedemos a la descarga de información usando el comando GET del paquete httr (v1.4.0). El contenido se descarga en formato JSON (JavaScript Object Notation), y es posteriormente procesado con dplyr (v0.8.3) y rlist (v0.4.6.1) hasta obtener un objeto data.frame para su fácil manipulación en R.

Mostramos dos ejemplos donde descargamos datos de las estadísticas vitales de defunciones perinatales precoces y tardías según semanas de gestación, y las cifras de población en España más recientes, por edad y provincia. Además, acompañamos la descarga de la información con gráficos realizados con ggplot2 (v3.2.0) para facilitar la interpretación y visualización de la información descargada.

Finalmente, mostramos la utilidad de la aplicación para la descarga de grandes volúmenes de información (Big Data), realizando descarga automática de más de seis millones de filas (población por edad simple por secciones censales de los años 2010, 2011 y 2012), haciendo uso de 156 URLs diferentes (52 provincias, 3 años). El tiempo aproximado de descarga fue de 75 horas en un ordenador con 8Gb de RAM.

Financiación: Instituto de Salud Carlos III (FIS PI18/01593), Subprograma de Vigilancia Epidemiológica del Cáncer (VICA) del CIBER de Epidemiología y Salud Pública (CIBERESP), EU-FEDER

Tratamiento de datos acelerométricos con `GENUDutils`.

Jorge Luis Ojeda Cabrera (Universidad de Zaragoza), Jorge Luis Ojeda Cabrera, Jorge Marin-Puyalto, Jose Antonio Casajus

La acelerometría se ha convertido en una herramienta indispensable a la hora de estudiar la actividad física de las personas de manera objetiva. Tanto en el ámbito deportivo como en el de la salud, y en particular en gerontología, el uso de acelerómetros de cintura y de muñeca es hoy en día una práctica habitual.

Dada la complejidad y tamaño de los datos que de proporcionan estos dispositivos (un registro semanal con una frecuencia de 30 Hz genera archivos de varios cientos de megabytes), su uso directo por parte de especialistas, terapeutas o para la asistencia sanitaria es imposible, siendo necesario su preprocesamiento.

En esta presentación se muestra el paquete `GENUDutils` cuyo principal propósito es el de dar soporte a dicho preprocesamiento y tratamiento
de los datos acelerométricos de manera sencilla y ágil, permitiendo el análisis simultáneo de varios registros acelerométricos. El paquete proporciona por una parte informes individuales para cada participante y por otra un resumen de resultados global, de forma que puedan ser analizados y estudiados por el personal competente.

El paquete proporciona utilidades básicas para leer datos acelerométricos crudos y, en base a ellos, desarrollar gráficos y calcular estadísticos y medidas de uso habitual en acelerometría. Estas utilidades u otras que se crean pertinentes se pueden incluir como código `R` en plantillas `Rmarkdown` según su necesidad. Utilizando estas plantillas personalizadas, el paquete `GENUDutils` genera tanto informes personalizados de cada registro individual como resúmenes globales de las mismas. La información necesaria para desarrollar todo el proceso se suministra al paquete mediante un pequeño interfaz gráfico, simplificando así su utilización.

Ajuste de modelos predictivos de series temporales para estimar los pedidos de un producto de consumo energético en una ciudad.

Carlos Pérez González (Universidad de La Laguna), Yanira Hernández Fernández; Marcos Colebrook Santamaría; José Luis Roda García; Sergio Díaz Martín; Jorge Antonio Herrera Alonso

Resumen: Las compañías distribuidoras de distintos productos de consumo energético (combustible, gas, etc..) necesitan resolver a diario problemas relacionados con la distribución de dichos productos hasta los diferentes puntos de suministro. El siguiente trabajo analiza la distribución de uno de dichos productos de acuerdo a los pedidos diarios realizados por los clientes en una ciudad durante el año 2018. De este modo, se obtienen estimaciones del consumo a corto plazo con el fin de ajustar el volumen de pedidos y la optimización de las rutas de transporte. En la predicción del número de pedidos de los clientes se han utilizado diversas librerías de R en las que utilizan diferentes modelos de series temporales como Holt Winters, ARIMA estacional y Prophet (Facebook). Estas librerías permiten realizar predicciones mediante el ajuste de tendencias no lineales a la estacionalidad anual, semanal y diaria, y consideran los efectos de las vacaciones y periodos festivos. Además son modelos robustos ante cambios en la tendencia y la falta de datos y tienen en cuenta, además, la presencia de observaciones atípicas. En Canarias, la compañía Distribuidora Industrial, S.A. (DISA) gestiona, junto a diferentes agencias de reparto, la distribución del gas de acuerdo a los pedidos que reciben cada día. A partir de los datos proporcionados por la compañía para el reparto en la ciudad de Las Palmas de Gran Canaria durante el año 2018, se lleva a cabo un estudio de la distribución de los pedidos realizados por los clientes. De este modo, se obtienen predicciones del consumo a corto plazo con el fin de ajustar el volumen de pedidos y la optimización de las rutas de transporte. El presente trabajo aborda la predicción del número de pedidos de gas utilizando diversas librerías de R que permiten ajustar diferentes modelos de series temporales (Holt Winters, ARIMA estacional y Prophet (Facebook). Estas librerías permiten realizar predicciones mediante el ajuste de tendencias no lineales a la estacionalidad anual, semanal y diaria, así como los efectos de las vacaciones. Además son modelos robustos ante cambios en la tendencia y la falta de datos y tienen en cuenta, además, la presencia de obsevaciones atípicas.

Sábado, 16 de noviembre

09:30 - 10:50: Premio Joven III: Aplicaciones en Medicina y Veterinaria

Aula/espacio: Salón de actos; Moderador/Responsable: Emilio López Cano

Modelos mixtos multinivel: Una aplicación de la librería lme4 para estimar el percentil de peso fetal en embarazos gemelares.

Rocío Aznar Gimeno (ITAINNOVA (Instituto Tecnológico de Aragón)), Luis Mariano Esteban (Escuela Politécnica La Almunia. Universidad de Zaragoza); Ricardo Savirón (Departamento de Obstreticia y Ginecología. Hospital General de Villarba); Gerardo Sanz (Departamento de Métodos Estocásticos. Universidad de Zaragoza)

La tasa de embarazos múltiples ha aumentado en los últimos años debido, principalmente, al aumento de la tecnología de reproducción asistida. El comportamiento de estos embarazos difiere al de los embarazos únicos y el uso de tablas de referencia de pesos desarrolladas en embarazos únicos para embarazos múltiples puede ocasionar un equivocado alto porcentaje de bajos pesos, lo que puede derivar en un posterior número de intervenciones innecesarias. Es por ello que existe necesidad clínica de desarrollar modelos específicos que permitan construir tablas de referencia de percentiles de peso para embarazos gemelares que permitan llevar un control del peso fetal a lo largo de la edad gestacional. El objetivo de nuestro estudio ha sido desarrollar modelos de crecimiento fetal para gestaciones gemelares según la corionicidad placentaria, a partir de los datos recogidos en el Hospital Miguel Servet de Zaragoza entre los años 2012 y 2017. Además, se han comparado con el rendimiento de otros modelos de crecimiento gemelares, construidos en diversas poblaciones (no

españolas y con características distintas), a la hora de predecir percentiles de peso pequeños y altos sobre nuestra población, para estimar así la conveniencia de sus usos en la práctica clínica.

El peso fetal es un dato longitudinal estimado mediante el uso de ultrasonido a lo largo de la edad gestacional. En particular, para los embarazos de gemelos, estos datos consisten en medidas repetidas para cada uno y ambos fetos de la misma madre. Esta jerarquía intrínseca de datos viola el supuesto estadístico de independencia. Para poder modelar la relación entre la edad gestacional y el peso estimado de estas medidas correlacionadas se han aplicado modelos lineales con efectos mixtos multinivel, capaces de resolver esta no-independencia.

Todo el estudio estadístico del trabajo fue desarrollado en el software R. Especialmente, se hizo uso de la librería lme4, con un gran potencial en el marco de los modelos con efectos mixtos y, en particular, de la función lmer que permitió el ajuste de estos modelos para el caso lineal. También se utilizó el paquete rms y, en particular, la función rcs para el uso de splines cúbicos restrictivos que permitió modelar la relación no lineal que se produce entre el peso fetal y la edad gestacional, así como para la estimación de los percentiles de peso para cada semana de gestación.

En conclusión, gracias al uso de diferentes funcionalidades que proporciona el software R, se ha podido conseguir desarrollar un modelo que prediga el peso de una forma ajustada para una determinada población de Aragón de embarazos gemelares y así construir tablas de referencia de percentiles de peso que permitirán tener un control más personalizado del peso fetal para este tipo de embarazos. Además, el uso del paquete shiny de R permitirá poner en producción estos resultados para su uso interactivo en la práctica clínica.

Derivación de patrones de dieta en niños de 4 y 8 años usando análisis de componentes principales con el paquete FactoMineR.

Eva María Navarrete Muñoz (CIBERESP-UMH), Alejandro Martínez-Moya(HGUE-FISABIO-UMH), Sandra Gonzalez-Palacios (ISABIAL-UMH); Desirée Valera-Gran (UMH-ISABIAL); Leyre Notario-Barandiaran(ISABIAL-UMH); Alejandro Oncina-Canovas (UMH-ISABIAL); Silvia Fernández-Barres (ISGlobal-UPF-CIBERESP); Jesús Ibarluzea (CIBERESP); Loreto Santa Marina (CIBERESP); Martine Vrijheid (ISGlobal, UPF, CIBERESP); Maribel Casas (ISGlobal, UPF, CIBERESP); Guillermo Fernández-Tardón (Universidad de Oviedo); Isolina Riaño (CIBERESP-ISPA - Universidad de Oviedo); Ana Espluges (UV-FISABIO-CIBERESP); Carmen Iñiguez (Dto Estadística e I.O. Universitat de València-Ciberesp); Manuela Garcia de la Hera (CIBERESP-ISABIAL-UMH); Jesús Vioque (CIBERESP-ISABIAL-UMH), Eva María Navarrete-Muñoz (CIBERESP-ISABIAL-UMH)

Los patrones dietéticos, derivados utilizando análisis de componentes principales, permiten evaluar la dieta de una forma global para cada individuo. El objetivo fue derivar patrones dietéticos en niños de 4 y 8 años en el estudio de cohortes prospectivo INMA. Se analizaron 1914 y 1765 niños a los 4 y 8 años respectivamente, con información dietética recogida mediante cuestionarios de frecuencia de alimentos (CFA) validados. La ingesta de alimentos del CFA se agrupó y estandarizó para 18 grupos de alimentos en g/día. La derivación de los patrones se llevó a cabo partiendo de estos grupos de alimentos normalizados con el paquete de R FactoMineR. No se usó ningún método automático para determinar el número de patrones, sino que se limitó al número mínimo cuya suma de la variabilidad fuera de al menos el 30% y con auto-valores mayores de 1.30. Para el etiquetado de los patrones se usó como criterio que las cargas factoriales de los grupos de alimentos fueran superiores $|0.3|$. A los 4 años, se seleccionaron los 3 patrones que explicaban una mayor variabilidad, un 12, 9.7 y 8.7% respectivamente. El primer patrón etiquetado como patrón Mediterráneo se caracterizó

por cargas superiores 0.3 en pescado y mariscos, verduras, legumbres, frutas y patatas. El segundo patrón se caracterizó por tener altas cargas para lácteo semi y derivados, carnes rojas y procesadas, fast food, dulces y reposterías y bebidas azucaradas y se etiquetó como patrón Occidental. El tercer patrón tuvo altas cargas para los lácteos enteros y derivados y lácteos semi y derivados y fue etiquetado como patrón Lácteo. A los 8 años, los 3 primeros patrones explicaban 11.2, 10.6 y 8.3% de la variabilidad respectivamente. Se observó que las cargas factoriales superiores a 0.3 fueron similares a las de los 4 años por lo que los patrones se etiquetaron de igual forma. El paquete de R FactorMineR permite derivar convenientemente patrones de dieta en estudios poblacionales para explorar el efecto de la dieta global en la salud de los niños.Financiación: Instituto de Salud Carlos III (PI18/0082) y Fundació la Marató de TV3 (201622-10)

Algoritmos de Machine Learning en detección de mastitis en vacas del trópico alto Colombiano.

Edimer David Jaramillo (Solla), Óscar Múnera Bedoya (Nutrisolla); Alexander Balbin Feria (Nutrisolla); Óliver Restrepo Rojas (Nutrisolla); Carolina Mesa Pineda (Nutrisolla)

La mastitis representa una problemática común en bovinos lecheros de todo el mundo, presentando inflamación de la glándula mamaria y tejido de la ubre. Constituida como una de las patologías más costosas en producción láctea. Se presenta en la mayoría de casos como subclínica y detectarla en estadíos tempranos, cuando carece de síntomas, es de vital importancia para disminuir pérdidas económicas, restringir uso de antibióticos y evitar transmisión a otros animales. Diagnosticar la patología a tiempo sugiere un reto para productores y profesionales del sector pecuario; por tal motivo, el objetivo del trabajo fue evaluar el desempeño de diferentes algoritmos de Machine Learning (ML) en la clasificación de vacas positivas o negativas a mastitis. Fueron utilizados registros productivos, reproductivos y ambientales para entrenar los modelos. Se probaron 10 algoritmos de clasificación supervisada, entre los cuales estaban Support-Vector-Machine, Red Neuronal, K-Nearest-Neighbor, entre otros. Se utilizaron conjuntos de entrenamiento y prueba, para medir el desempeño del modelo a través de Validación Cruzada y Bootstrapping. La elección de los mejores predictores se llevó a cabo a través de eliminación recursiva con Random Forest y Bootstrapping, basado en la precisión de clasificación. El proceso de análisis de datos, desde la depuración hasta la modelación, fue llevado a cabo con el sistema estadístico R, haciendo uso de bibliotecas destacadas en Data Science como Tidyverse y caret. El recuento de células somáticas, niveles de grasa, proteína, sólidos totales, edad del animal y promedio de producción de leche, fueron considerados como predictores de alta importancia. Las mejores clasificaciones se consiguieron con K-Nearest-Neighbors, Red Neuronal y Support-Vector-Machine, con precisiones de testing de 84.2%. Estos resultados permiten afirmar que los algoritmos de ML se constituyen como herramientas matemáticas y computacionales que pueden ser usadas en campo para generar alertas tempranas de patologías subclínicas en la industria láctea.

Desarrollo de un sistema de control de calidad y auditoría previo a la implementación del cribado de preeclampsia del primer trimestre en España. Análisis secundario del estudio PREVAL.

Valeria Rolle Sóñora (ISPA), Diana Cuenca Gómez (Servicio de Obstetricia y Ginecología. Hospital Universitario de Torrejón, Madrid); Catalina de Paco Matallana (Servicio de Obstetricia y Ginecología. Hospital Clínico Universitario Virgen de la Arrixaca, Murcia); Nuria Valiño Calviño (Servicio de Obstetricia y Ginecología. Hospital Universitario de A Coruña, Coruña); Begoña Adiego Burgos (Servicio de Obstetricia y Ginecología. Hospital Universitario Fundación de Alcorcón, Madrid); Rocío Revello Álvarez ((Servicio de Obstetricia y Ginecología. Hospital Universitario

Quirón Pozuelo, Madrid); Belén Santacruz Martín (Servicio de Obstetricia y Ginecología. Hospital Universitario de Torrejón, Madrid); María del Mar Gil Mira (Servicio de Obstetricia y Ginecología. Hospital Universitario de Torrejón, Madrid)

El estudio PREVAL pretende validar el cribado de PE en el primer trimestre en 20.000 gestantes en Cataluña, Coruña, Madrid y Murcia. Se consideró imprescindible el desarrollo de un sistema de control de calidad y auditoría que garantizase la calidad del cribado. Para ello se desarrollaron una serie de scripts en el software R que, con periodicidad mensual, generasen de manera automática representaciones gráficas de los parámetros auditables: tensión araterial mediam (TAM), índice de pulsatilidad de las arterias uterinas (IPAut), factor de crecimiento placentario (PlGF y proteína plasmática asociada al embarazo A. Desde el 15 de septiembre de 2017 hasta el 1 de marzo de 2019 se reclutaron 9185 pacientes. Se observó una una desviación de los múltiplos de la mediana (MoM) de la TAM a una mediana de 0.95 y del PlGF y la PAPP-A a 0.85 y 1.08 MoM. Se comprobó la correcta estimación del IPAut. Tras la retroalimentación a los investigadores se corrigió la TAM a una mediana de 0.99 MoM y el PlGF y la PAPP-A permanecieron estables, por lo que actualmente se está elaborando una ecuación correctora.

09:30 - 10:50: Economía y Empresa

Aula/espacio: Fdez. Huerta; Moderador/Responsable: Pedro Concejero

Online Atlas, o cómo identificar áreas idóneas para invertir en proyectos renovables.

Rubén Moreno (DNV GL), Circe Triviño (DNV GL)

Presentamos una herramienta online interactiva de visualización y análisis de información geográfica, especialmente diseñada para ayudar a inversores y desarrolladores de proyectos renovables (eólicos y solares) en las etapas iniciales de prospección y evaluación de proyectos, para que puedan explorar los mejores sitios para invertir y desarrollar, e identificar ágilmente los riesgos que podrían frenar sus proyectos. Se trata de una herramienta web, creada fundamentalmente utilizando R y su amplio catálogo de librerías y componentes. Muestra el potencial del ecosistema R para construir aplicaciones comerciales complejas, cubriendo prácticamente todas las necesidades. Desde el uso de Shiny como plataforma base para la aplicación web, al manejo, visualización y análisis de datos espaciales, pasando por aspectos más técnicos como la integración con bases de datos, la autenticación de usuarios o la gestión de cálculos costosos de forma asíncrona. El objetivo es ofrecer una herramienta potente, y al mismo tiempo muy fácil de manejar incluso para usuarios sin experiencia en Sistemas de Información Geográfica, SIG, que reúna toda la información y funcionalidades clave en una única aplicación web. No solo proporciona un completo catálogo de capas y fenómenos geográficos relevantes que afectan la localización de un proyecto renovable. También incluye información estadística actualizada sobre la evolución de distintos indicadores. Además, la aplicación también permite al usuario realizar consultas, configurando y combinando una amplia variedad de filtros de una manera muy intuitiva para descubrir, analizar y seleccionar al instante las ubicaciones más idóneas.

Análisis y evaluación de la sensibilidad al precio utilizando el paquete partykit.

Jorge Martín Arevalillo (UNED)

En este trabajo se lleva a cabo una revisión de la metodología del particionamiento recursivo basado en modelos y se muestra su relevancia en el acercamiento al problema del análisis y evaluación de la

sensibilidad al precio. Se presenta un enfoque que permite aplicar la metodología a dicho problema utilizando las prestaciones del paquete partykit y se señala su potencial en la identificación de grupos de clientes cuya sensibilidad al precio presenta un patrón diferencial. La metodología se aplica a un caso de negocio con datos reales de un producto crediticio. Los resultados obtenidos proporcionan una segmentación que permite clasificar los clientes en función de su sensibilidad al precio, identificando perfiles con una baja sensibilidad frente a otros cuya sensibilidad es alta o muy alta. Se examinan las implicaciones que tiene este hallazgo en el diseño de estrategias personalizadas de fijación de precios, haciendo un énfasis especial en el alto valor estratégico de los resultados como inputs de negocio para abordar el problema de la fijación óptima del precio.

Multivariate forecast of Ecuadorian financial indexes using Gaussian DBNs with Bnlearn.

David Quesada (UPM), Gabriel Valverde; Pedro Larrañaga; Concha Bielza

El estudio y evolución de las entidades financieras de un pa´ıs en ocasiones se realiza sobre distintos ´ındices evaluados de manera independiente y sin tener en cuenta las relaciones que se dan entre las entidades financieras. Para modelar estas dependencias condicionadas, proponemos el uso de Redes Bayesianas Dinámicas para predecir la evolución conjunta de todos los´ındices financieros medidos sobre los bancos ecuatorianos. Hemos utilizado las funcionalidades de redes bayesianas gaussianas que encontramos en bnlearn para desarrollar su versión dinámica.

09:30 - 10:50: Metodología y paquetes

Aula/espacio: Ricardo Marín; Moderador/Responsable: Román Mínguez

Compositional software: coda.base, zCompositions and CoDaPack.

Jose Antonio Martin Fernandez (Universitat de Girona (UdG)), M. Comas-Cufí (Universitat de Girona); D. Re (Universitat de Girona); S. Thió-Henestrosa (Universitat de Girona); J. Palarea-Albaladejo (BIOSS); J.A. Martín-Fernández (Universitat de Girona)

Compositional data (CoDa) are commonly defined as multivariate data vectors of strictly positive components whose sum is constant (e.g., 1, 100, 1, 10^6). CoDa analysis methods are nowadays having a notable impact in varied fields such as economy, biology, medicine, and engineering; and their application requires the use of specialised software packages. The R packages "coda.base" and "zCompositions", developed within the CoDa-Research group at the University of Girona, include a number of functions to make CoDa analysis accessible to the general scientific community. They however share the standard characteristics of R packages based on command-line instructions, which prevents practitioners with no basic acquaintance with programming from using them. CoDaPack is a user-friendly Java based application with basic capabilities and high quality graphical outputs with interactive features. Recently, several R functionalities have been integrated under the hood to work within CoDaPack, so that users can use them and obtain the same outputs through menus. CoDaPack makes use of its own separated version of R. The user can have another version of R installed on the system. This connection with R makes the most usual multivariate techniques such as regression, MANOVA, cluster and discriminant analysis seamlessly available in CoDaPack. Moreover, a hidden developer menu allows to actually execute R scripts interacting directly with the CoDaPack, generating outputs including textual R outputs, graphical R outputs and new entire data frames or single columns which can be added to the current CoDaPack worksheet.

IndTestPP: un paquete R para analizar la dependencia entre procesos puntuales.

Ana C. Cebrián (Universidad de Zaragoza)

IndTestPP es un paquete R que proporciona un amplio conjunto de tests y herramientas estadísticas para analizar la dependencia entre procesos puntuales, condicionalmente a su estructura marginal. El paquete tiene implementadas tres familias de tests para analizar la independencia entre procesos puntuales: la primera está basada en la distribución Poisson de los procesos, la segunda ofrece tests basados en un bootstrap paramétrico, y la tercera tests basados en una aproximación de tipo Lotwick-Silverman. Estas familias cubren una amplia variedad de situaciones: procesos de Poisson, procesos homogéneos y no homogéneos, procesos puntuales sin hipótesis adicionales, etc. El paquete también proporciona medidas de dependencia y funciones para generar diferentes tipos de procesos puntuales dependientes e independientes. Estas funciones son útiles por ejemplo para la implementación de herramientas de inferencia basadas en métodos computacionales.

Las propiedades y utilización de IndtestPP se muestra con el análisis de la dependencia entre la ocurrencia de sucesos extremos de calor en tres localidades españolas, Barcelona, Zaragoza y Huesca. El objetivo es establecer si la ocurrencia de extremos de calor en las tres localidades es independiente, o si existe dependencia, identificar los factores que la explican.

spsur: an R package for spatial seemingly unrelated regressions.

Román Mínguez Salido (Universidad de Castilla-La Mancha), Fernando A. López (Universidad Politécnica de Cartagena); Jesús Mur (Universidad de Zaragoza)

In recent decades, several methodological improvements have been suggested to specify, estimate and validate Seemingly Unrelated Regression (SUR) models in a spatial framework. These new procedures allow us testing for the presence of spatial dependence, and estimate spatial models using different algorithms. The new package spsur allows for the estimation and inference of the most popular spatial econometric models by maximum likelihood or instrumental variable procedures. Additional functions allowing for the estimation of the so-called spatial impacts (direct, indirect and total impacts) and to simulate spatial SUR DGP's are included. The package also includes well-known examples in the applied literature on spatial data.

09:30 - 10:50: Medicina y Genética

Aula/espacio: Salón de grados; Moderador/Responsable: Luis Mariano Esteban

genehummus: Automatización computacional para el estudio de familias génicas basadas en la organización de los dominios conservados.

Jose Die Ramón (Universidad de Córdoba), Jose V Die (National Center for Biotechnology Information; Departamento Genética, Universidad de Córdoba) ; Ben Busby (National Center for Biotechnology Information)

Esta propuesta se enmarca dentro del área de conocimiento de la Genética, Genómica y Bioinformática. En caso de aceptarse, daríamos una visión general del uso que hacemos de genehummus en la Escuela de Ingenieros Agrónomos de la Universidad de Córdoba (UCO). genehummus es un paquete escrito en R que hemos desarrollado en los Institutos Nacionales de Salud de los EEUU (Bethesda, Maryland) para la identificación automatizada de familias génicas. La aparición reciente

de secuencias completas del genoma de especies con interés agronómico es una oportunidad excelente para la identificación a escala genómica de familias génicas mediante el uso de herramientas bioinformáticas. La identificación de los miembros de una familia génica determinada sirve de información base para entender la evolución de esas secuencias o la función que desarrollan en un organismo o tejido. Ofreceremos casos prácticos, donde a partir de genehummus identificamos, por ejemplo, genes implicados en la tolerancia a sequía.
El esquema general de la presentación sería el siguiente: en primer lugar, el problema que representa la anotación "manual" de secuencias genéticas; a continuación explicaríamos la ventaja que ofrece un sistema automatizado; después hablaríamos brevemente de la API del Centro Nacional para la Información Biotecnológica de los EEUU (NCBI); nos detendríamos con más detalle en la forma en la que implementamos la librería (descripción de dependencias, funciones, y warnings); finalmente haríamos un breve comentario sobre nuestra experiencia durante el proceso de envío del paquete a CRAN. Aunque genehummus se ha desarrollado tomando el modelo de plantas, anticipamos que puede aplicarse sobre cualquier especie.

El código de genehummus está completamente disponible en nuestro repositorio : https://github.com/NCBI-Hackathons/GeneHummus donde hay también documentación detallada sobre su uso. Por último, genehummus puede instalarse desde CRAN : https://CRAN.R-project.org/package=geneHummus .

La regresión de Poisson con varianza robusta vs logística en R: El caso de la dieta Mediterránea y alteración del procesamiento sensorial en el proyecto InProS.

Eva María Navarrete Muñoz (CIBERESP-Universidad Miguel Hernández), Eva María Navarrete-Muñoz; Paula Fernández-Pires, Miriam Hurtado-Pomares; Paula Peral-Gómez; Iris Juarez-Leal; Cristina Espinosa-Sempere; Alicia Sánchez-Perez y Desirée Valera-Gran (afiliación de todas Universidad Miguel Hernández)

El proyecto InProS es un estudio transversal llevado a cabo en la provincia de Alicante con niños de 3 a 7 años cuyo propósito fue explorar factores asociados, tales como adherencia a la dieta Mediterránea (DM), a alteraciones del procesamiento sensorial (PS). En el presente trabajo examinamos las diferencias en la asociación múltiple entre adherencia a la DM y las alteraciones del PS, usando regresión logística para estimar odds ratios o regresión de Poisson de varianza robusta basado en el método del sándwich de Huber. El PS se evaluó usando la adaptación española del Short Sensory Profile (SSP) y su alteración se definió como SSP total <155. La adherencia a la DM se midió con el KIDMED y se clasificó a los niños en adherencia baja, media y alta. La prevalencia de alteración del PS fue de 29.8%. Comparado con los de baja adherencia, los de adherencia media tuvieron RP= 0.77; IC 95%: 0.54-1.12 vs OR= 0.55; IC95% :0.32-0.93, y los de adherencia alta tuvieron RP= 0.77; IC 95%: 0.54-1.12 vs OR=0.44; IC 95%: 0.21; 0.95. Los resultados con ambos métodos de regresión muestran una reducción de las alteraciones del PS con una mayor adherencia a la DM; sin embargo, la asociación para OR es significativa y la de la RP no. Ante el posible sesgo de sobreestimación, cuando la prevalencia es superior al 20%, la RP resulta ser un mejor estimador que la OR.

Machine Learning en Salud: predicción de riesgo cardiovascular a 10 años (R-Studio y Rapid Miner).

Juan José Beunza Nuin (Universidad Europea de Madrid), Juan José Beunza (Univ. Europea de Madrid, Director grupo Machine Learning Salud-UEM); Enrique Puertas (Univ. Europea de Madrid, Director Master Big Data)

El machine learning acaba de aterrizar en el mundo sanitario. Ya hay más de 23 productos en proceso de aprobación por la FDA (Food & Drug Administration), y todo parece indicar que su crecimiento seguirá siendo exponencial. En España disponemos ya de herramientas válidas para la aplicación de algoritmos de machine learning a nuestros datos clínicos. R es una de las herramientas clave, además de Python y las nubes. Inicialmente pretendemos hacer un breve repaso de la realidad del desarrollo de aplicaciones de machine learning aplicados al entorno sanitario y de que cabe esperar de las oportunidades comerciales futuras. El corazón del taller consistirá en explicar los detalles del desarrollo de un algoritmo de predicción de riesgo cardiovascular (infarto de miocardio) a 10 años que hemos desarrollado en el grupo de trabajo Machine Learning Salud-UEM. Explicaremos primero su desarrollo en R-Studio, con los obstáculos que nos hemos ido encontrando por el camino, tanto en la obtención de datos clínicos locales como en los resultados de aplicar distintos algoritmos y arquitecturas, y como los hemos ido resolviendo. Luego explicaremos brevemente cómo hemos desarrollado el mismo algoritmo con el software comercial Rapid Miner (gratuito para entornos académicos), y haremos una comparativa de nuestra experiencia con los dos lenguajes. Terminaremos abriendo un debate de opiniones y de posible creación de networking entre los interesados en el campo.

fcaR, un paquete R para manipulación de implicaciones difusas: Diseño de un sistema de recomendación para diagnóstico médico.

Domingo López-Rodríguez (Universidad de Málaga), Ángel Mora Bonilla, Domingo López Rodríguez

En este trabajo, presentaremos el paquete fcaR, cuyo objetivo es la extracción y manipulación de implicaciones (reglas de asociación con confianza 1), tanto difusas como no difusas. La intención de los autores es subir este paquete a CRAN lo antes posible.

El paquete está basado en técnicas formales: Formal Concept Analysis y en la Lógica de Simplificaciones en las que los autores del paquete estamos trabajando.

Se explicarán las funciones principales y métodos que se han desarrollado en el paquete y cómo, gracias a ellos, y a diferencia del uso habitual de paquete similares como arules, somos capaces de razonar con las implicaciones y deducir nuevo conocimiento muy interesante.

Como funciones implementadas en el nuevo paquete destacamos la eliminación de redundancia y el cálculo de cierres de ítems que pueden ser clave, por ejemplo, para el diseño de sistemas de recomendación.

Acabaremos con una aplicación práctica de este paquete al diseño de un sistema de recomendación con un dataset relacionado con el diagnóstico médico.

11:20 - 12:20: Ponente invitado: Jo-Fai Chow

Aula/espacio: Salón de actos; Moderador/Responsable: Carlos Ortega

Automatic and explainable machine learning in R.

Jo-Fai Chow (H2O)

Resumen pendiente

Patrocinador Platino

Deloitte.

Patrocinadores Oro

UNED

R consortium

cajamar CAJA RURAL

innova-tsn

orange™

Patrocinadores Plata

R Studio

idealista/data

Data Centric

H_2O.ai

DSLAB DATA SCIENCE LABORATORY

U-tad CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

PIPERLAB

Patrocinadores Bronce

Tau Analytics

mf MÁXIMA formación

UMUR

Apoyo Institucional

REPSOL

SEIO