I Congreso & XII Jornadas de Usuarios de R

Universidad de Córdoba

23-25 noviembre 2022

Talleres

Jueves 15-17h (NAT1)

Vitaminando el análisis de las variables nominales con R en Ciencias de la Vida

Mercedes Ovejero Bruna. *Unidad de Biometría, Sermes CRO, Madrid, España / Departamento de Metodología y Psicobiología, Universidad Complutense de Madrid, Madrid, España*

Elena Rosa-Pérez. *Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España*

David Salgado. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

Las variables nominales permiten indicar cualidades que se expresan con palabras y no se ordenan según un criterio de jerarquía. El análisis de estas variables es una herramienta útil y necesaria especialmente en Ciencias Sociales y de la Salud, pudiéndose extender a cualquier área de la Ciencia. Las técnicas más conocidas permiten estudiar la asociación entre estas variables y obtener conclusiones que abren las puertas a análisis más complejos. No obstante, existen otras alternativas que también aportan importante información analítica relacionada con el estudio de cambios longitudinales, comparación de grupos e incluso el análisis de la calidad de modelos avanzados. El objetivo del presente taller es proporcionar una visión del análisis de las variables nominales más allá del estudio de su asociación, todo ello desde R mediante un enfoque participativo y eminentemente aplicado. La estructura del taller es la siguiente: 1. Concepto y ejemplos de variables nominales. 2. Análisis descriptivo y visual de las variables nominales: de la distribución de frecuencias a la tabla de contingencia. 3. Asociación entre variables nominales: a. Coeficiente ji-cuadrado. b. Coeficientes de contingencia, Phi y V de Cramer. c. Coeficiente lambda de Goodman-Kruskal. 4. Análisis de las frecuencias marginales: a. La prueba de McNemar. La prueba de Bowker. b. La prueba Q de Cochran. 5. Análisis de las odds ratio y el riesgo relativo. 6. Introducción al análisis de la matriz de confusión. 7. Conclusiones y líneas de aprendizaje avanzadas. Se proporcionará el código de R necesario para llevar a cabo el taller y seguir las explicaciones pertinentes.

#ciencias #vida

Jueves 15-17h (NAT2)

Mi primer blog con Quarto (y Rmarkdown)

Pedro J. Pérez. *Departamento de Análisis Económico, Universitat de València, Valencia, España*

Gemma Pérez-López. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

José-Luis Zafra-Gómez. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

Un punto fuerte de R es el ecosistema Rmarkdown que facilita los análisis reproducibles y la creación de contenidos en múltiples formatos. La mayoría de usuarios de R utilizan Rmarkdown en sus análisis, pero menos lo utilizan para crear blogs, ya estén estos orientados a la docencia o a la divulgación, a pesar de que, si conoces Rmarkdown, la barrera de entada es baja.

La creación y gestión de un blog tiene beneficios evidentes como la visibilidad e interrelación con los lectores; además, escribir es una forma excelente de aprender. Escribir un post sobre un análisis o tema R sobre el que estás aprendiendo facilita profundizar, recordar y retomarlo luego con mayor facilidad; además, escribir posts es divertido. Vamos, que en mi opinión todo el mundo debería tener un blog.

Recientemente ha aparecido Quarto, un programa creado por RStudio que se anuncia como la segunda generación de Rmarkdown. A pesar de que Rmarkdown no va a desaparecer, es probable que Quarto acabé sustituyéndolo.

El taller comenzará, tras exponer brevemente los beneficios de tener un blog, explicando qué es Quarto, ventajas, similitudes y diferencias con Rmarkdown. La segunda parte se dedicará a la creación del blog, repasando la estructura de ficheros y principales opciones de personalización. La tercera se simulará la escritura de posts para practicar la dinámica de gestión del blog y creación de contenidos con Quarto. Finalmente se procederá a publicarlo, seguramente a través de Github Pages.

#creación #blogs #rmarkdown #quarto

Jueves 15-17h (NAT3)

R a alta velocidad y fácil “a la dplyr”.

Carlos Ortega Fernández. *QUALITYEXCELLENCE SL, MADRID, ESPAÑA*

El objetivo de la presentación es dar a conocer las altas capacidades que ofrece “R” para el procesado de datos grandes y a alta velocididad.

Se presentará data.table, pero usando “tidytable” que permite trabajar de forma “a la dplyr” evitando así, las complejidades de la sintaxis de data.table.

Durante la charla se trabajará sobre un conjunto de datos grande demostrando las capacidades y facilidad del uso de “tidytable”.

En el desarrollo del taller, se presentarán datos de forma gráfica usando ggplot y otros paquetes asociados que simplifican el uso incluso de ggplot.

Destacar que por incluir un elemento importante de dinamismo y frescura, el taller tendrá una importance componente interactiva y dinámica. El enfoque del análisis será el que se acuerde entre los asistentes al taller.

#caso #práctico #usando #conjunto #datos #grande

Comunicaciones Oral

Ciencias sociales y economía

Eficiencia para un desarrollo sostenible: análisis de los servicios públicos con el software R

Pedro José Martínez Córdoba. *Departamento de Administración y Economía de la Empresa, Universidad de Salamanca, Salamanca, España*

Elena Rosa-Pérez. *Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España*

David Salgado. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

Para alcanzar un desarrollo sostenible en la gestión de los servicios públicos, analizamos la eficiencia y sus factores determinantes con el software R. Las múltiples ventajas que ofrece R para calcular, comparar e interpretar los resultados permite realizar todo el proceso metodológico con el mismo software. El objetivo de este trabajo consiste en mostrar la implementación de R en el ámbito de la contabilidad y auditoría del sector público. Ante la falta de una metodología clara y estándar para el cálculo de la eficiencia, utilizamos los tres métodos no paramétricos más aceptados por la literatura: el Data Envelopment Analysis (DEA); su versión no convexa, Free Disposable Hull (FDH); y el Order-m. Para calcular estos niveles de eficiencia con R utilizamos la librería nonparaeff. En línea con estos métodos no paramétricos encontramos el índice Malmquist, considerada la herramienta más adecuada para medir la evolución de la eficiencia según los cambios tecnológicos producidos. Para calcular este índice utilizamos la librería deaR. Tras obtener los niveles de eficiencia, estimamos diferentes modelos de regresión según los objetivos de investigación. De esta forma, utilizamos modelos de regresión truncada con la librería truncreg; modelos de regresión logística (Generalized Linear Models); o modelos de mínimos cuadrados ordinarios (Ordinary Least Squares). Otras librerías como readxl; pastecs; psych; o corrplot nos facilitan los análisis previos. Las librerías presentadas forman parte de la metodología implementada para el cálculo de la eficiencia en los servicios públicos. La incorporación de R a las investigaciones en contabilidad y auditoría del sector público permiten avanzar en la calidad de los trabajos. En el futuro, conforme al espíritu que caracteriza a la comunidad R-Hispano, se debe continuar actualizando las librerías existentes, así como realizar otras que permitan la elaboración de bases de datos más flexibles, rápidas y dinámicas.

#ciencias #económicas

Ciencias sociales y economía: Jueves 10-11h (CiO2)

Construcción de Modelos de Credit Scoring con R

Francisco Jesús Rodríguez Aragón. *Oney, Madrid, España*

Gemma Pérez-López. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

José-Luis Zafra-Gómez. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

Los modelos de credit scoring quizás sean los más utilizados a día de hoy por el entorno bancario y los que suelen ser más fácilmente aceptado por parte de la regulación bancaria. Sin embargo a pesar de su uso y de la gran cantidad de ingreso que genera su aplicación efectiva resultan prácticamente desconocidos por la comunidad de DS que suelen optar a menudo por opciones mucho más complejas, donde se pierde la interpretabilidad y la comprensión total del proceso que va desde la toma del dato, hasta la modelización final, con todas sus fases intermedias

#modelización #estadística

UP IN THE AIR: Predicción Del Número De Pasajeros En Puente Aéreo Madrid-Barcelona

Julián Rojo. *Universidad de Extremadura, Innova-tsn*

El caso práctico que vamos a presentar versa sobre la predicción del número de pasajeros en los vuelos del puente aéreo Madrid-Barcelona. La complejidad de realizar esta predicción radica en que la ocupación de estos vuelos es extremadamente variable, puesto que además de ocuparse con pasajeros que han realizado una reserva previa del billete, las plazas se complementan con viajeros que realizan cambios de un vuelo a otro, al tratarse de tarifas flexibles, o de aquellos que compran el billete poco antes de que el vuelo despegue. Además, se requieren diferentes horizontes de predicción suponiendo un cálculo de unas 8.400 predicciones diarias que corresponden a los vuelos anuales de la compañía.

Para realizar estas predicciones contamos con un histórico de 4 años con variables relativas al vuelo, pasajeros en vuelos equivalentes, así como las reservas previas de cada vuelo. El tratamiento de esta información nos permite generar un tablón con 175 variables. A través de varios modelos ensamblados disponibles en las librerías de R (KNN, K-Means, Regresión lineal y SVM) se estiman diferentes predicciones que sirven de entrada al modelo final, un XGBoost, que es el que determina la predicción definitiva. Además, hay una serie de vuelos clasificados como críticos, en los que la predicción no puede estar por debajo del volumen final real de pasajeros, por lo que se requiere una sobreestimación. En estos se realiza una Regresión cuantílica quedándonos con el cuantil 90 como valor. Gracias a este proceso se ha conseguido reducir el error de la predicción hasta en un 60% en el horizonte de 120 días, reduciendo además los tiempos de dedicación, pasando de tener a una persona dedicada en exclusiva a esta tarea, a tener un proceso que genera y envía estas predicciones en 4 horas de manera autónoma.

Proyecto desarrollado por Innova-tsn para IBERIA

#analítica #predictiva #machine #learning #sector #transportes #turismo

Estimaciones avanzadas en los Índices de Cifras de Negocios de la Industria asistida por algoritmos de Machine Learning

Sandra Barragán. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

Begoña Vega. *Innova-tsn*

María Neira. *Innova-tsn*

Ángela Díaz. *Innova-tsn*

Las estadísticas económicas coyunturales son las estadísticas más rápidamente producidas y difundidas para mostrar las tendencias en la economía europea. Su gran ventaja descansa en su disponibilidad poco después de los correspondientes períodos de referencia (mensuales y cuatrimestrales). Debido a la pandemia por COVID-19, las oficinas de estadística han sido requeridas para mejorar la oportunidad (timeliness) de estos productos acortando más la diferencia entre los períodos de publicación y de referencia. Inevitablemente, esto conduce a un delicado equilibrio entre la oportunidad y la acuracidad.

Presentamos un proceso de producción que permitiría obtener los Índices de Cifras de Negocios de la Industria (ICN) desde los primeros días que los datos son recogidos por la oficina de estadística, produciendo índices adelantados a partir de los 10 días del comienzo de la recogida de los microdatos. Esto supone una reducción notable respecto al intervalo de publicación actual (51 días). Para ello, seguimos la filosofía de los estimadores GREG y de Sanguiao-Zhang para proponer, específicamente para un diseño muestral por cut-off, un estimador proyectivo-predictivo. En esta combinación, se utilizan directamente los microdatos recogidos y se predicen los aún por recoger.

Las predicciones se realizan mediante técnicas de machine learning construyendo regresores a partir de microdatos y de paradatos de períodos de referencia pasados y en curso de la misma operación estadística (ICN) al igual que de otras operaciones coyunturales relacionadas. El algoritmo empleado ha sido un boosting gradiente ligero (lgbm) sobre árboles de regresión.

La mejora de la dimensión de la calidad de oportunidad (timeliness) es obvia pero se observa un fuerte compromiso con la determinación de la acuracidad. Usamos técnicas convencionales combinadas con la hipótesis de intercambiabilidad (exchangeability) de las unidades estadísticas para calcular el error cuadrático medio, cifra de referencia para evaluar cuán precisas son las estimaciones avanzadas respecto de las finales.

#producción #estadísticas #oficiales

R, machine learning e innovación en la producción de estadísticas oficiales: algunas reflexiones con casos de uso

David Salgado. *Depto. Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España; Depto. Estadística e Investigación Operativa, Universidad Complutense de Madrid, España*

Andrés Vinueza. *Departamento de Ciencia de Datos, LOGIKARESEARCH Cía. Ltda, Quito Ecuador*

Jorge Sosa. *Departamento de Matemática, Escuela Politécnica Nacional, Quito Ecuador*

Defendemos que R está altamente adecuado y adaptado para la innovación en la producción de estadísticas oficiales. Esto es especialmente evidente en la incorporación de las técnicas de aprendizaje automático (machine learning) en el conjunto de métodos estadísticos oficiales. Proporcionamos tres ejemplos ilustrativos, a saber (i) el incremento de la eficiencia por coste en la depuración de datos, (ii) la mejora de la oportunidad (timeliness) mediante la producción de estimaciones avanzadas con un ejercicio de imputación en masa y (iii) la desagregación temporal de estimaciones basadas en diseños muestrales mediante la computación de pesos de muestreo para periodos temporales de referencia más cortos (de trimestrales a mensuales). Todos estos casos de estudio se han desarrollado completamente en R. Presentamos resultados con datos de operaciones estadísticas oficiales reales subrayando la adecuación de R para la innovación de la metodología estadística en una oficina de estadística. R permite a los expertos en diversas materias, metodólogos, e informático producir de modo colaborativo herramientas de software estadístico fiables de modo muy rápido. Esto trae a la vida productos viables mínimos (minimum viable products) de manera muy dinámica, que pueden ser posteriormente desarrollados y desplegados en producción. Afirmamos e ilustramos con ejemplos que R es especialmente sobresaliente en acortar y acelerar el trayecto desde las ideas metodológicas hasta los prototipos con datos reales preparando el camino para una combinación rigurosa de las metodologías en poblaciones finitas y el aprendizaje estadístico (statistical learning). Estas ideas conectant directamente con iniciativas en el Sistema Estadístico Europeo con la creación de un grupo de trabajo interno sobre Open Source Software y la impresionante lista de herramientas OSS clasificadas de acuerdo con el GSBPM (https://github.com/SNStatComp/awesome-official-statistics-software (https://github.com/SNStatComp/awesome-official-statistics-software)).

#producción #estadísticas #oficiales

Ciencias sociales y economía: Viernes 11-12h (CiO6)

Big data al servicio de las ciudades

Miguel Flores. *Grupo MODES, SIGTI, Departamento de Matem'atica, Escuela Politécnica Nacional, Quito Ecuador*

Las tecnologías de información y las comunicaciones son el medio para construir ciudades inteligentes (Smart cities) que sean eficientes, sostenibles y amigables con el ambiente. Los pilares o módulos de un sistema (geoportal) sugeridos para administrar una Smart city son: información social, gestión territorial, pulso de la ciudad, gestión de tráfico y Smart buisness.

El cantón Manta ubicado en el Ecuador ha centrado sus esfuerzos en convertirse en la primera Smart city del país, para lograrlo ha iniciado con el primer pilar para el control social. Este módulo se ha desarrollado utilizando R para el cálculo de indicadores de accesibilidad y desigualdad social y el paquete shiny para la interfaz gráfica (geoportal).

Esta herramienta informática permitirá a las autoridades del cantón Manta: analizar las condiciones de vida de los ciudadanos, e identificar los sectores priorizados que requieren gestionar política pública emergente.

#ciencias #sociales #humanidades

Utilizacion de R en los procesos de preservacion digital de una biblioteca

Fernando Martinez de Guzmán. *Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid*

Francisco Angel Guerrero Vivas. *Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid, España*

Luis Martínez Uribe. *Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid, España*

Comienza a ser habitual encontrar bibliotecas y centros de información que incorporan el uso del lenguaje de programación R en distintas áreas de trabajo con datos digitales. Una de estas áreas es la preservación digital, una labor cuyo esfuerzo se centra en garantizar que la información digital de valor siga siendo accesible y utilizable.

El Datalab de la Biblioteca/Centro de apoyo a la investigación de la Fundación Juan March es el responsable del sistema de preservación digital, que dota a la Fundación de políticas, infraestructuras y procesos para preservar los contenidos digitales que se generan a diario.

La utilización de R en los procesos de preservación es fundamental, tanto para la creación de colecciones, como para la ingesta de datos y objetos digitales en dichas colecciones, así como para la monitorización de los distintos elementos del sistema, y la corrección de posibles errores. Ciertos desarrollos permiten además la sincronización automática entre los procesos de producción de contenido y la preservación de los objetos digitales creados.

Esta charla presentará la herramienta web de monitorización del sistema de preservación desarrollada en R-Shiny, así como diversos desarrollos en R para los procesos de ingesta, comprobación de datos y corrección de errores. Al mismo tiempo se mostrarán desarrollos en R para la sincronización automática entre los sistemas de creación de contenido y el flujo de los objetos digitales creados hacia el sistema de preservación.

#bibliotecas #preservación #digital

OnomasticDiversity: un paquete para cuantificar regiones desde la onomástica

María José Ginzo Villamayor. *Santiago de Compostela*

Sandra Barragán. *Depto. Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

El análisis de las relaciones entre los apellidos y las características genéticas se remonta a finales del siglo XIX, con el estudio de los apellidos para calcular la probabilidad de los matrimonios entre primos hermanos en Gran Bretaña. De hecho, es probable que los individuos que comparten apellidos específicos de una localidad también compartan una serie de características lingüísticas, genéticas, históricas y sociales, así como una ascendencia común. Normalmente, estos análisis se basan en medidas de isonimia. La isonimia se refiere a la posesión del mismo apellido.

Este trabajo se centrará en la introducción de nuevos métodos estadísticos para el tratamiento y modelización de datos en geolingüística, concretamente, en los apellidos de Galicia. El objetivo principal es la modelización de patrones espaciales y espacio-temporales de apellidos en esta región. Las diferentes líneas de investigación en el contexto onomástico no han tenido en cuenta la dimensión espacial y espacio-temporal de la evolución de los apellidos. Además, se presentará el paquete de R, OnomasticDiversity, donde se recoge la implementación de la mayor parte de estas técnicas.

#matemáticas

Docencia y programación R

estadistica: un paquete de R para estadistica descriptiva e inferencial.

Vicente Coll Serrano. *Departamento de Economía Aplicad, Universidad de Valencia, Valencia, España*

Elena Rosa-Pérez. *Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España*

David Salgado. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

En los campos de la economía y la empresa, la aplicación del análisis estadístico con datos reales debería ser un requisito básico para comprender los procesos y dinámicas económicas reales y para que los estudiantes puedan pensar como profesionales. La Guidelines for Undergraduate Programs in Statistical Science de la American Statistical Association hace referencia, entre otros, (1) al aumento de la importancia de la ciencia de los datos, queriendo destacar la relevancia de las habilidades informáticas para los estudiantes, y (2) las aplicaciones reales. Respecto a esta última, el análisis de datos reales requiere que los estudiantes se familiaricen con el software de análisis estadístico profesional, para aprender a acceder y manipular los datos de diversas maneras orientadas a la resolución de problemas reales. Hoy en día el software de análisis estadístico de referencia es R. Sin embargo, el manejo de R presenta ciertas barreras de entrada porque su uso requiere el aprendizaje de una sintaxis. Por esta razón, hemos programado una librería de R, a la que hemos denominado “estadística”, que trata de minimizar estas barreras de entrada. En esta comunicación presentaremos algunas de las funcionalidades de esta librería: funciones implementadas, documentación de las funciones, video-tutoriales disponibles para facilitar el autoaprendizaje, etc.

#matemáticas

Presentación de la audioguia: “Introducción a la Ciencia de Datos con R”Ciencia de Datos con R. Preparación de los datos y Análisis no supervisado”

Juana María Alonso Revenga. *Departamento de Estadística y Ciencia de los datos, Facultad de Estudios Estadísticos, Universidad Complutense, Madrid, España*

Gemma Pérez-López. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

José-Luis Zafra-Gómez. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

La importancia del análisis de datos en nuestros días es indiscutible. Para realizar un correcto análisis de datos y poder extraer toda la información que contienen, ahora que disponemos de grandes cantidades de datos, se han desarrollado técnicas que unen la metodología estadística clásica y las herramientas informáticas de programación, lenguaje y algoritmos, dando lugar a lo que podemos llamar Ciencia de Datos. Presentamos un libro en formato audioguía para introducirnos de una forma clara y práctica en las técnicas más utilizadas en Ciencia de Datos utilizando la metodología de resolución de casos prácticos con R. Hacemos especial hincapie en la interpretación correcta de los resultados obtenidos en la aplicación de las técnicas presentadas porque en esta parte del proceso es en donde se presentan las mayores dudas y se pueden cometer los mayores errores. Se presenta un recorrido de once temas (divididos en dos volúmenes) en donde comenzamos por explicar los conceptos fundamentales de la Ciencia de Datos y el software elegido. Hemos elegido el lenguaje R con el entorno Rstudio porque es uno de los lenguajes más utilizados en Ciencia de Datos. El volumen I contiene los temas de preparación del conjunto de datos y análisis no supervisado y en el volumen II presentaremos las técnicas de predicción y clasificación. Todos los temas tienen incluidos videos explicativos de las técnicas presentadas, mostradas siempre mediante ejemplos. Además, los datos utilizados en estos ejemplos están disponibles en la plataforma de Ingebook para que el alumno pueda utilizarlos como guía en su trabajo.
Esperamos que este material sea de utilidad para aquellos que se inician en esta ciencia tanto en los múltiples grados de Ciencia o Ingeniería de Datos, como en los másteres de Big Data, Inteligencia de Negocio y otras muchas áreas relacionadas con el análisis de datos.

#ciencias #informáticas #ingeniería

Aprendiendo a programar en R en las Ciencias de la Tierra

Pablo González Moreno. *Departamento de Ingeniería Forestal, Universidad de Córdoba*

Las Ciencias de la Tierra tienen como objetivo común el estudio de nuestro planeta, incluyendo la biosfera, su evolución, e interacciones. Esta área de conocimiento incluye disciplinas tan variadas como las ciencias forestales y agronómicas, biológicas o ambientales, compartiendo un interés común por generar herramientas que ayuden a comprender las causas y patrones de los procesos naturales y facilitar una explotación sostenible de los recursos naturales en sistemas naturales y agropecuarios. Es en este ámbito dónde la programación, especialmente en R, ha surgido como una herramienta indispensable para estas disciplinas al facilitar la aplicación y reproducibilidad de técnicas de análisis de datos avanzadas a un volumen de datos que no es posible asimilar con instrumentos menos potentes. Por este motivo, los estudios superiores en ciencias de la tierra están empezando a incorporar habilidades de programación en sus curricula. Sin embargo, no es una tarea fácil por el propio perfil de los estudiantes, con un nivel tecnológico variado y un interés especial en la aplicación más que en la propia programación. En esta comunicación, aportamos la experiencia de diversas asignaturas de grado y máster en ciencias de la tierra en las que se ha incorporado la programación en R, bien como actividad práctica o como elemento principal de la asignatura. En primer lugar mostramos las distintas aproximaciones que se han aplicado a nivel de grado y máster, poniendo ejemplos concretos de actividades y prácticas realizadas. Finalmente, desde la experiencia del profesorado, reflexionamos sobre las experiencias docentes, su papel facilitando el aprendizaje, así como sus limitaciones en la adquisición de los conocimientos y destrezas en programación.

#medioambiente #ciencias #geográficas

Los paquetes van a CRAN

Lluís Revilla Sancho. *Laboratorio de enfermedad inflamatoria intestinal, IDIBAPS, Barcelona, España*

Begoña Vega. *Innova-tsn*

María Neira. *Innova-tsn*

Ángela Díaz. *Innova-tsn*

Por lo general, compartir nuestro trabajo con la comunidad R significa enviar un paquete a un archivo (CRAN, Bioconductor u otros). CRAN es el mayor repositorio de paquetes de R que viene aceptado por defecto por R. Pero, ¿Qué hay que hacer para escribir un paquete, que CRAN lo acepte y se mantenga en CRAN?

Que un paquete se mantenga en CRAN depende de la calidad del paquete. Esto se debe a que hay que pasar un proceso de revisión. Si el buen paquete sigue las reglas y tiene una calidad de acuerdo con sus criterios, se durará. Primero, hay un chequeo inicial automático; segundo, una revisión manual más profunda del código. Luego, si las sugerencias se aplican o se responden correctamente, el paquete se incluye en el archivo.

En cada paso se utilizan algunas reglas y criterios para decidir si el paquete avanza o no. Comprender lo que dicen estas reglas, los problemas comunes y los comentarios de los revisores ayudarán a evitar enviar un paquete para que sea rechazado. Reducir la fricción entre compartir nuestro trabajo, proporcionar paquetes útiles a la comunidad y minimizar el tiempo y los esfuerzos de los revisores.

A partir de los datos históricos veremos el proceso habitual, el tiempo de espera hasta su inclusión, el número de revisiones habituales antes de ser aceptados y el porcentaje de éxito. También haremos un recorrido histórico de los paquetes de CRAN: tiempo de duración de una versión en CRAN, relación entre versiones y dependencias y el número de paquetes nuevos habituales. Para ver qué características tienen que cumplir nuestro paquete para ser incluido y que otros usuarios pueden usar nuestro código con garantías de calidad.

#desarrollo #paquetes

Estadística y análisis de datos

optedr: un paquete de diseño óptimo de experimentos

Carlos de la Calle Arroyo. *Instituto de Ciencia de los Datos e Inteligencia Artificial, Universidad de Navarra, Pamplona, España*

Elena Rosa-Pérez. *Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España*

David Salgado. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

A menudo en el diseño óptimo de experimentos los esfuerzos se enfocan bien en generar diseños óptimos para problemas particulares o familias de problemas, bien en el desarrollo de algoritmos y procedimientos generales para encontrar dichos diseños. No obstante, en la práctica a menudo sucede que un diseño es óptimo para un criterio, pero aun así no es adecuado para su uso real. Los experimentadores pueden tener necesidades concretas o restricciones, preferencias para ciertos puntos experimentales, necesidades estadísticas, etc. En dichos casos, ante la imposibilidad de implementar directamente el diseño óptimo, se puede optar por usar el diseño óptimo como un benchmark, o bien aumentar o modificar el diseño para adecuarlo a las preferencias del práctico. El paquete optedr permite el cálculo de diseños óptimos para modelos no lineales con una variable independiente, dentro de la familia exponencial, para diferentes criterios. Este se ha desarrollado con un enfoque aplicado, con una interfaz y manejo sencillos para generar los diseños. Además del cálculo de diseños óptimos, el paquete permite comparar diseños propuestos por el usuario con otros diseños óptimos, para el criterio de optimalidad de estos. Asimismo, implementa una metodología para D-aumentar diseños de una manera informativa, controlando la eficiencia del diseño resultante. Por último, como en el paquete se trabaja con diseños aproximados, se ha implementado un algoritmo de redondeo para transformar los diseños óptimos aproximados y diseños aumentados en diseños exactos, listos para su uso por el experimentador.

#matemáticas

ExactTree: Un paquete de R para obtener árboles de decisión optimizados globalmente.

Juan Claramunt González. *Methodology and Statistics, Leiden University, Leiden, The Netherlands*

Gemma Pérez-López. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

José-Luis Zafra-Gómez. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

Los árboles de decisión, tales como los árboles de clasificación y regresión (CART por sus siglas en inglés) se construyen, en general, por medio de particiones binarias recursivas. Su objetivo predecir categorías o clases (árboles de clasificación) o valores (árboles de regresión) de la mejor manera posible utilizando particiones binarias en base a las variables independientes.

La mayoría de los algoritmos utilizados para construir estos árboles comienzan con un nodo inicial con todos los sujetos. Dado ese nodo, buscan la variable y el valor que optimicen la división de los sujetos en base a un criterio local que minimice la impureza (por ejemplo, minimizar la suma de los cuadrados de los residuos). Dados la variable y el valor obtenidos, unos sujetos son asignados a un nodo u otro del siguiente nivel del árbol. Este proceso se realiza de manera recursiva hasta construir todo el árbol. El potencial problema de estos algoritmos es que pueden obtener un mínimo local. Por ello, aparecen nuevos algoritmos que buscan el mínimo global de la función que define la impureza, como “evolutionary tres” (evtree) o ExactTree. Por ejemplo, el algoritmo a presentar, ExactTree, es un algoritmo de programación dinámica exacta que trata de optimizar la completa estructura del árbol con respecto a una función objetivo global. Esto es, ExactTree no optimiza cada división, sino que optimiza toda la estructura del árbol en su conjunto para obtener el mínimo global de la función objetivo.

Hemos realizado un estudio comparando distintos algoritmos para la construcción de árboles de decisión. Los resultados muestran que la precisión de las predicciones es similar, pero la estabilidad de ExactTree es superior. En la presentación, se mostrarán los resultados del estudio y se hará una demostración de cómo utilizar el paquete de R, ExactTree.

#matemáticas

EAT: un paquete de R para la estimación de funciones de producción a través de técnicas de aprendizaje automático

Víctor Javier España Roch. *Instituto Centro de Investigación Operativa (CIO), Universidad Miguel Hernández, Elche, España*

EAT es un paquete de R que contiene funciones para la estimación de fronteras de producción y medidas de eficiencia técnica utilizando técnicas de aprendizaje automático no paramétrico: Árboles de regresión y Bosques Aleatorios. El paquete implementa los principales algoritmos asociados con una nueva técnica introducida para estimar la eficiencia de un conjunto de unidades de toma de decisiones en Economía e Ingeniería, llamados Árboles de Análisis de Eficiencia. Además, incorpora un amplio abanico de funciones orientadas a facilitar la interpretación de los resultados al usuario como representaciones gráficas o rankings de importancias de variables.

#matemáticas

Estadística y análisis de datos: Jueves 12-14h (EsO4)

Cómo facturamos millones de dólares con nuestros paneles web construidos dinámicamente en R y Shiny

David Durey. *Departamento de Frogtek Analytics, Frogtek, Huesca, España*

Begoña Vega. *Innova-tsn*

María Neira. *Innova-tsn*

Ángela Díaz. *Innova-tsn*

Soy el CPO de Frogtek Analytics. Nos dedicamos a informatizar tiendas tradicionales en países en vías de desarrollo y obtener, por primera vez para esos lugares, datos en tiempo real de todos los movimientos (compras y ventas) de productos de consumo. Luego, también nosotros mismos, agregamos y limpiamos esa información y calculamos las tendencias del mercado, que es una información tremendamente valiosa para los grandes fabricantes de productos. Por eso somos una empresa social (informatizamos y ayudamos a las pequeñas tiendas de forma casi gratuita) y lucrativa (les cobramos por la información agregada a las grandes compañías). La etapa final del ciclo de datos de toda nuestra infraestructura está programada en R y nuestros paneles web con la información, en Shiny. Hemos desarrollado, además, la capacidad de modularizar al extremo (mediante parámetros entendibles por cualquier persona que no sepa nada de programación) la generación automática y completa de las webs, de forma dinámica y desatendida. Estamos listos para encajar fácilmente 1 nuevo cliente o 200, cada uno con su web personalizada y customizada según sus necesidades. Me encantaría enseñarle a la comunidad cómo lo hemos conseguido.

#ciencias #informáticas #ingeniería

Estadística y análisis de datos: Jueves 12-14h (EsO5)

Muestrear no es pecado

Jose Luis Cañadas Reche. *Orange Spain, Madrid, España*

Andrés Vinueza. *Departamento de Ciencia de Datos, LOGIKARESEARCH Cía. Ltda, Quito Ecuador*

Jorge Sosa. *Departamento de Matemática, Escuela Politécnica Nacional, Quito Ecuador*

En el mundo del big data actual se olvidan con frecuencia los conceptos básicos de la estadística, tales como el muestreo, el resumen de los datos vía frecuencias o incluso el análisis bayesiano. En esta comunicación cuento un pequeño ejemplo como utilizar R en entorno de Big Data para tratar con grandes volúmenes de datos, y también como el muestreo y otras técnicas se pueden aplicar en estos entornos.

#big #data #divulgación

Análisis de condiciones necesarias (NCA) como complemento de medidas tradicionales de tamaño del efecto.

Ana Martina Greco. *Estudis de Dret i Ciència Política, Universitat Oberta de Catalunya, Barcelona, España*

Actualmente está en discusión el uso de coeficientes para medir el tamaño del efecto en ciencias sociales (p. ej., r de Pearson, d de Cohen). Se ha argumentado que los puntos de cortes sugeridos no siempre son aplicables a las ciencias sociales y que son coeficientes de difícil interpretación o aplicación práctica. Recientemente se ha generado un interés creciente en combinar estas medidas con otras técnicas que permitan una interpretación adecuada para generar intervenciones o acciones propias de las ciencias sociales. Entre estas técnicas, se encuentra el análisis de condiciones necesarias (necessary condition analysis, NCA, Dul et al., 2016). Este método permite cuantificar cuanta cantidad de la variable independiente X (condición) es necesaria para permitir que ocurra la variable dependiente Y (resultado), a través de un tamaño del efecto d que se calcula en base a la parta vacía de del gráfico de correlación. Además, permite calcular la significación estadística de este efecto, así como obtener puntos de corte de cada condición para acceder a un mínimo de la variable dependiente, y puntos de ineficiencia tanto para el resultado como para las condiciones que se quieran testear. Se muestra una prueba del uso de este método en una muestra de 235 estudiantes de psicología, analizando la condiciones necesarias del rendimiento previo, horas de estudio y rasgo de personalidad responsable. También se acompaña esta técnica con la tradicional medida del tamaño del efecto de Pearson. Finalmente, se interpretan los resultados de cada uno de los análisis, resaltando la posible utilidad y aplicabilidad de la información obtenida del NCA para la toma de decisión en el contexto del ejemplo.

#ciencias #sociales #humanidades

pspatreg: an R package for semiparametric modelling of spatio-temporal data

Román Mínguez Salido. *Universidad de Castilla-La Mancha*

Francisco Angel Guerrero Vivas. *Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid, España*

Luis Martínez Uribe. *Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid, España*

We propose a semiparametric P-Spline model to deal with spatial panel data. This model includes a non-parametric spatio-temporal trend, a spatial lag of the dependent variable, and a time series autoregressive noise. Specifically, we consider a spatiotemporal ANOVA model, disaggregating the trend into spatial and temporal main effects, as well as second- and third-order interactions between them. Algorithms based on spatial anisotropic penalties are used to estimate all the parameters in a closed form without the need for multidimensional optimization. We implement an R package, named pspatreg, which allows to estimate and make inference for this type of models.

#econometría #espacial #espacio #temporal

2 shiny apps para el ocio y el negocio

Leonardo Hansa. -

Sandra Barragán. *Depto. Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

Presento dos apps de Shiny. Una es una interfaz para un juego habitual entre grupos de amigos. Se juega introduciendo en un recipiente nombres escritos en papelitos. Luego, por equipos, hay que describir uno a uno esos nombres para que los miembros del equipo adivinen el personaje que hay escrito. Replico ese esquema de juego en Shiny.

El segundo Shiny es una propuesta de negocio. Se trataría de una aplicación en la que los usuarios son inquilinos que podrían valorar las casas en las que han vivido de alquiler. Así, cuando una persona esté interesada en alquilar una casa, podrá consultar las críticas que tiene esa casa. El Shiny es solo un prototipo de cómo sería la usabilidad de la app.

#desarrollo #herramientas #web

El paquete “multiColl” para detectar multicolinealidad preocupante

Román Salmerón Gómez. *Departamento de Métodos Cuantitativos para la Economía y la Empresa, Universidad de Granada, Granada, España*

Jaime Ballesteros de la Vega. *Unidad de Biometría, Sermes CRO, Madrid, España*

María Villar Navales. *Universidad Complutense de Madrid, Madrid, España*

Valeria Samira Samanamud Taus. *Universidad Complutense de Madrid, Madrid, España*

La regresión lineal múltiple es una herramienta econométrica que permite establecer relaciones entre un conjunto de variables, conocidas como independientes, con la variable de interés, denominada dependiente, y es ampliamente usada en distintos ámbitos como el biosanitario, medioambiental, económico o deportivo. Uno de los posibles problemas que se puede presentar al aplicar esta técnica es el de multicolinealidad (relaciones lineales entre las variables independientes de la regresión). En este caso, pueden aparecer distintas consecuencias, siendo una de ellas el incumplimiento del ceteris paribus. Es decir, no sería posible aislar el efecto de las variables independientes sobre la dependiente, siendo este uno de los objetivos principales al aplicar esta técnica. Por tal motivo, es importante detectar cuándo la presencia de multicolinealidad debe de preocupar al investigador. Con este objetivo, en este caso, se presenta el paquete “multiColl” de R, el cual es comparado con otros paquetes existentes en R que también se centran en detectar este problema.

#ciencias #económicas

Estadística y análisis de datos: Jueves 12-14h (EsO10)

Un paseo por la Luna con R

Francisco Jesús Rodríguez Aragón. *Socio R-Hispano, Madrid, España*

Fuensanta Arnaldos-García. *Departamento de Métodos Cuantitativos para la Economía y la Empresa. Universidad de Murcia*

M. Teresa Díaz-Delfa. *Departamento de Métodos Cuantitativos para la Economía y la Empresa. Universidad de Murcia*

En esta exposición vamos a ver con R como es posible aplicar un mapa lunar sobre una fotografía de esta, debidamente graduada, para señalizar distintos accidentes geográficos y ayudar a los astrónomos en sus observaciones, además mediante shiny se tendrá accesibilidad desde el móvil

#astronomía

El test BDS: definición, usos y aplicaciones al análisis de series temporales con R

Lorenzo Escot. *Facultad de Estudios Estadísticos, Universidad Complutense de Madrid, España*

El test BDS (Brock, W.A; Scheinkman, J.A.; Dechert, W.D; & LeBaron, B, 1996) es un potente test que surgió de la aplicación de la teoría del Caos al análisis de series temporales que sirve para contrastar la hipótesis nula de i.i.d., esto es, que los elementos de la serie están independiente e idénticamente distribuidos. La gran capacidad de este test BDS para detectar dependencias temporales (lineales y no lineales) ha motivado su incorporación en la diagnosis de los modelos de series temporales (librería `tseries`) para contrastar la hipótesis nula de ruido blanco. La potencia de este test para detectar dependencia (o independencia) temporal en una serie nace del hecho de que, a diferencia del tradicional correlograma (acf y pacf), no se basa en el concepto de correlación temporal sino en el concepto de integral de correlación. Esta integral de correlación mide la dependencia espacial de la órbita reconstruida en el espacio de fases con la serie temporal. Y por ello el correcto uso del test BDS, y de los diferentes parámetros que es necesario fijar para su estimación (retardo, dimensión de inserción y radio de cercanía espacial), requiere entender cómo y porqué es posible reconstruir la trayectoria del sistema generador de la serie temporal (aunque dicho sistema sea desconocido). En esta, comunicación presentaremos este test BDS y cuál es la mejor estrategia o regla para su correcta aplicación. Presentaremos también algunos ejemplos de uso, tanto para la diagnosis de los residuos de modelos ARIMA, como otras aplicaciones menos conocidas como test para contrastar la existencia de cambios estructurales en el modelo generador de la serie cuando este test BDS se aplica de manera recursiva.

#ciencias #económicas

Entendiendo el gráfico CUSUM mediante simulación

Emilio López Cano. *Centro de Investigación para las Tecnologías Inteligentes de la Información y sus Aplicaciones (CETINIA), Universidad Rey Juan Carlos*

Luis Martinez-Uribe. *Fundación Juan March*

Carlos Prieto. *Universidad de Salamanca*

David Barrios. *Universidad de Salamanca*

Cristina Calvo. *Universidad de Salamanca*

Los gráficos de control son herramientas muy potentes dentro de las técnicas de Control Estadístico de Procesos. Los gráficos de control de Shewhart son los más sencillos de crear y muy sencillos de interpretar. En particular el par de gráficos de la media y el rango pueden ser utilizados por “dueños del proceso” sin una formación avanzada en Estadística (aunque siempre ayuda). Sin embargo, otros gráficos más especializados no son tan sencillos de interpretar, y requieren más entrenamiento. El gráfico CUSUM (de cumulative sums, sumas acumuladas) es uno de estos gráficos. Su ventaja con respecto a los gráficos de Shewhart es que pueden detectar pequeños cambios en los procesos, que pasarían desapercibidos en los gráficos de Shewhart. En esencia, es similar a cualquier gráfico de control. Se fija una línea central y unos límites de control superior e inferior entre los cuales se van monitorizando los estadísticos calculados de cada subgrupo. Y aquí es donde se complica la interpretación, porque los valores que se representan no tienen una interpretación inmediata. Además, se representan dos series de valores: las diferencias positivas con los valores centrales, y las diferencias negativas. La línea central suele ser el cero, pero el diseño de los límites de control no es trivial, y a menudo se realiza probando distintas posibilidades hasta que obtener el diseño que mejor sirve a los propósitos del control de procesos. En este trabajo se presenta un enfoque basado en la simulación que permite entender qué efectos van a producir en los gráficos de control cambios en el proceso. Se presentarán también algunos usos innovadores de este enfoque para utilizar los gráficos CUSUM en el diseño y monitorización de estudios comparativos, incluyendo un ejemplo real en estudios nutricionales de producción porcina.

#ciencias #informáticas #ingeniería

IoTimpute: una aplicación shiny para la imputación de valores faltantes para el Internet de las Cosas

Aurora González Vidal. *Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, España e ITI-CERTH, Salónica, Grecia.*

Aida Calviño Martinez. *Departamento de Estadística y Ciencia de los datos, Facultad de Estudios Estadísticos, Universidad Complutense, Madrid, España*

IoTimpute: una aplicación shiny para la imputación de valores faltantes para el Internet de las Cosas

La imputación de valores faltantes en una serie o conjunto de datos es un viejo conocido. A día de hoy, los despliegues de Internet de las Cosas hacen que dispongamos de grandes cantidades de datos. El Internet de las cosas (IoT) permite la integración de sensores, actuadores y dispositivos de comunicación para aplicaciones en tiempo real. Pero para tomar decisiones basadas en datos en tiempo real se necesitan datos de buena calidad y, en parte, esto se relaciona con la ausencia de valores faltantes.

Existen una gran cantidad de paquetes y métodos estadísticos para la imputación de valores faltantes, sin embargo, nos hemos enfocado en aquellos capaces de adaptarse a las características de los datos del Internet de las Cosas: 1) Espacio-temporalidad 2) Calidad de los sensores

Para ello se han realizado experimentos con los métodos Bayesian Maximum Entropy (BME) y Probabilistic Matrix Factorization (PMF), llegando a la conclusión de que el primero supera al segundo en características y precisión. BME es un método de mapeo para la estimación espacio-temporal y PMF, método iterativo muy utilizado en los sistemas de recomendación.

Para la utilización de ambos modelos se ha desarrollado una aplicación shiny que permite considerar la calidad de los sensores (alta-baja), las coordenadas de los mismos (su proximidad) y con la que se pueden tanto imputar valores faltantes directamente como hacer pruebas con datos completos añadiendo un rango determinado de valores faltantes para comprobar cuáles son los parámetros más apropiados a seleccionar para un caso de uso en cuestión con los modelos programados.

#ciencias #informáticas #ingeniería

Implementación en R del método Kernel Weighting para la reducción de sesgos en muestras no probabilísticas

Jorge Luis Rueda Sánchez. *Departamento de Estadística e Investigación Operativa, Universidad de Granada, Granada, España*

Rosario Martínez-Verdú. *Departamento de Economía Aplicad, Universidad de Valencia, Valencia, España*

A causa de los grandes beneficios que se obtienen de los cuestionarios online y del Big Data, las muestras no probabilísticas han adquirido un gran valor tanto para la sociedad como para las empresas en la actualidad. Sin embargo, debido a su propia naturaleza presentan una gran cantidad de sesgos que pueden dar lugar a estimaciones imprecisas. Destaca especialmente el sesgo de voluntariedad, que será importante si existen diferencias significativas entre los individuos muestreados y los no muestreados.

En este trabajo estudiamos una nueva e innovadora técnica de reducción de sesgos de voluntariedad llamada Kernel Weighting (KW) que ha ofrecido resultados muy prometedores. Para comprobar su eficacia realizaremos un estudio de simulación comparándola con otra técnica ampliamente estudiada, y cuya eficacia ya ha sido contrastada, como es el Propensity Score Adjustment (PSA). En esta sesión se introducirá la implementación de estas técnicas en R mediante los paquetes NonProbEst y KWML, que nos ofrecen todo tipo de funciones para calcular estimaciones a partir de muestras no probabilísticas usando diferentes técnicas de reducción de sesgos.

`#matemáticas`

BGWBP: A new tool to teach branching processes

María del Pilar González Barquero. *Departamento de Matemáticas, Universidad de Extremadura, Badajoz, España*

Salvador Arenas-Castro. *Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba*

Francisco Javier Bonet García. *Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba*

Jose V. Die. *Departamento de Genética, Universidad de Córdoba*

Diego Nieto Lugilde. *Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba*

Francisco J Ruiz Gómez. *Departamento de Ingeniería Forestal, Universidad de Córdoba*

Branching processes describe the evolution of systems whose elements reproduce following probability laws, so that moving from one state of the system to another is done essentially convolutely. Due to the origin of the processes in a population dynamic context, systems are usually referred as “populations” and their elements as “individuals”. In the simplest branching model, known as Bienaymé-Galton-Watson process (BGWP), individuals are independent from each other and each of them generates a random number of descendants following a common probability law, after which it disappears from the population. This work consists of several functions created in the statistical software R and grouped in a library called “BGWBP”. The aim of these functions is to represent the extinction probability and the asymptotic or limiting behavior, analytically and graphically, of the three classes of BGWPs according to their offspring mean, critical, subcritical or supercritical. Moreover, several functions are also included to calculate the maximum likelihood and moment methods estimates for the main parameters of a BGWP.

#matemáticas

evolMap: mapas interactivos y evolutivos para la visualización de fuentes de datos en R

Carlos Prieto Sánchez. *Servicio de Bioinformatica, Nucleus, Universidad de Salamanca.*

evolMap es un nuevo paquete de R que permite representar la información de una base de datos sobre un mapa geográfico interactivo. Los datos se pueden representar sobre el mapa mediante marcadores, líneas o coropletas, que permiten adaptar su aspecto visual en función de la información de la base de datos. Los elementos representados pueden filtrarse o localizarse en el mapa mediante filtros interactivos y un cuadro de búsqueda. La aplicación también permite representar vínculos entre marcadores para realizar visualizaciones de redes sobre el mapa. Toda la información de la base de datos se puede consultar interactivamente en los paneles de información, tablas dinámicas y popups. evolMap no solo solventa la limitación del software existente en la visualización de bases de datos, también permite visualizar la evolución temporal o en periodos de los elementos añadidos. De modo que el mapa se puede reproducir como una visualización interactiva cambiante en el tiempo. La representación geográfica ha sido implementada sobre la librería Leaflet de JavaScript y su desarrollo ha sido inspirado en el paquete netCoin para la visualización y el manejo de datos interactivos. Las propiedades que hemos desarrollado convierten a evolMap en una plataforma única para explorar datos sobre un mapa geográfico y mostrar su evolución en el tiempo. evolMap está abierto para su uso en https://github.com/BioinfoUSAL/evolMap (https://github.com/BioinfoUSAL/evolMap) y se pueden visualizar algunos ejemplos en https://bioinfo.usal.es/evolMap (https://bioinfo.usal.es/evolMap).

#visualización #interdisciplinar

Paquete skd: contrastes de hipótesis basados en distancias kernel

Bojan Mihaljevic. *Universidad Politécnica de Madrid*

El problema de dos muestras o problema de homogeneidad es uno de los más estudiados en estadística por su interés en ciencia y tecnología. Destacan por su simplicidad e interpretabilidad los tests de homogeneidad basados en distancias o métricas. Junto al creciente interés en nuevos tipos de datos (alta dimensión, FDA, datos en variedades,…) han aparecido nuevas propuestas de métricas que permiten afrontar este problema. Un ejemplo de ello son las distancias kernel. En nuestra charla presentaremos el paquete skd que implementa el test de homogeneidad basado en la supremum kernel distance, nuestra propuesta para mejorar los tests basados en distancias kernel, además de varios otros tests basados en esta distancia.

`#estadística`

INLAMSM: Building on top of R-INLA to analyse multivariate spatial models

Francisco Palmí Perales. *Departamento de Economía Aplicada, Universitat de València, València, España*

Jesús López-Fidalgo. *Instituto de Ciencia de los Datos e Inteligencia Artificial, Universidad de Navarra, Pamplona, España*

Licesio J. Rodríguez-Aragón. *Escuela de Ingeniería Industrial y Aeroespacial de Toledo, Universidad de Castilla-La Mancha, Toledo, España*

Bearing in mind that multivariate procedures lead to high computation time procedures, different ideas have been developed to speed up fitting multivariate spatial models. Some authors prefer to handle this problems using approximate methods, such as the INLA method (Rue et. al. 2009) which is commonly used in order to avoid the high computation time of the MCMC models.

The goal of this work is to implement, using R-INLA environment, several multivariate areal models which were proposed for been computed using MCMC algorithms and then re-parametrized in order to be more computationally efficient. An R package, INLAMSM, which includes all the methods implemented, has been developed and it is available on CRAN. Specifically, this R package provides a collection of multivariate spatial models for analysing lattice data. The implemented models, which include different structures to model the variables' spatial variation and the between-variables variability, can be used (with R-INLA) for performing Bayesian inference. Therefore, fitting multivariate spatial models becomes faster and easier with INLAMSM.

Two different datasets have been used to exemplify the use of the package. The results and the computation time of the different implemented methodologies have been analysed for this two examples.

References

Botella-Rocamora, P., Martinez-Beneito, M.A., Banerjee, S. (2015). A Unifying Modeling Framework for Highly Multivariate Disease Mapping. Statistics in Medicine, 45 1548-1559.

Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology), 71(2), 319-392.

#estadística #aplicada

Medioambiente y ciencias geográficas

De los datos al bosque

Mª Ángeles Varo Martínez. *Departamento de Ingeniería Forestal, Universidad de Córdoba, Córdoba (España)*

Elena Rosa-Pérez. *Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España*

David Salgado. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

Las ciencias forestales, están viviendo una drástica transformación gracias a los grandes avances tecnológicos de las últimas décadas. En el contexto en el que estamos inmersos, el alto grado de innovación y el dinamismo científico en esta área de conocimiento nos han llevado desde la libreta y el lápiz a pie de árbol a la inteligencia artificial y a la automatización de procesos. R cuenta con numerosas librerías específicas que participan en las ciencias forestales actuales, por ejemplo, leaflet, mapview, raster o sf que permiten la visualización de mapas interactivos y la gestión de datos provenientes de inventarios forestales; getLandsat para la descarga de imágenes satelitales o lidR para el análisis de información de sensores remotos conocidos como LiDAR que posibilitan un análisis fisiológico del estado sanitario del bosque. El objetivo de la comunicación es presentar las técnicas que ofrece R que permiten una gestión forestal sostenible de última generación, desde un análisis multitemporal de los datos recogidos por distintos organismos, un estudio de la evolución del bosque a través de sensores remotos y una separación del monte en unidades homogéneas que faciliten su gestión. Para ello, se utilizarán como hilo conductor dos casos de estudio, por un lado, el monte de Pinar de Yunquera, con la presencia de una especie forestal emblemática como es el pinsapo y, por otro, el incendio ocurrido en 1993 en el interior del Parque Natural de la Sierra de Huétor.

#ciencias #forestales

RecordTest: Un paquete de R para detectar comportamientos no estacionarios en la ocurrencia de eventos récord

Jorge Castillo Mateo. *Departamento de Métodos Estadísticos, Universidad de Zaragoza, Zaragoza, España*

Gemma Pérez-López. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

José-Luis Zafra-Gómez. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

El estudio de comportamientos no estacionarios en los extremos es importante para analizar datos climáticos o medioambientales, entre otros. Como una alternativa a la teoría clásica de valores extremos, el estudio de eventos que baten nuevos récords resulta particularmente atractivo. El aumento medio de temperatura durante los últimos años está bien caracterizado y estudiado, sin embargo, a pesar de su presencia en los medios de comunicación, la caracterización del cambio ocurrido en los récords de temperatura sigue siendo un frente abierto. En este trabajo presentamos el paquete de R RecordTest disponible en CRAN (https://CRAN.R-project.org/package=RecordTest (https://CRAN.R-project.org/package=RecordTest)) y en GitHub (https://github.com/JorgeCastilloMateo/RecordTest (https://github.com/JorgeCastilloMateo/RecordTest)). Este paquete proporciona un marco para el análisis no paramétrico del comportamiento no estacionario en los extremos, basado en el análisis de récords. La idea principal de las herramientas desarrolladas corresponde a comprobar si los récords observados en los datos son compatibles con la distribución de la ocurrencia de récords bajo series estacionarias de variables aleatorias. Se proponen distintos contrastes de hipótesis para detectar tendencias o cambios de punto en la ocurrencia de récords basados en los récords superiores pero también en récords superiores e inferiores de la serie hacia delante y hacia detrás. El paquete también implementa todos los pasos necesarios en este tipo de análisis como la preparación de los datos, la identificación de los récords, herramientas exploratorias, y herramientas gráficas complementarias que hacen uso del paquete ggplot2. Finalmente, para ilustrar la gran variedad de herramientas incluidas en el paquete y la utilización de las mismas, se muestra un análisis detallado cuyo objetivo es el estudio del efecto del calentamiento global en los récords de una serie de temperatura máxima diaria de la Península Ibérica.

#medioambiente #ciencias #geográficas

GREENeR: Un paquete R para estimar y visualizar la presión de los nutrientes en las aguas superficiales

Angel Udias Moinelo. *Dpto. Ciencias de la Computación, Arquitectura de Computadores, Lenguajes y Sistemas Informáticos y Estadística e Investigación Operativa, Universidad Rey Juan Carlos, Mostoles, Madrid, Españ*

La contaminación por nutrientes afecta a las aguas dulces y costeras de todo el mundo. La planificación de acciones eficaces para mitigar los impactos de la contaminación requiere herramientas para evaluar los flujos de nutrientes liberados por las actividades humanas en las aguas, y las posibles reducciones mediante políticas de restauración. Cada vez se dispone de más datos sobre las emisiones y concentraciones de nutrientes en las estaciones de control, con una mejor resolución espacial y temporal. Los modelos conceptuales de cuenca hidrográfica aprovechan los datos disponibles, proporcionando una interpretación física de las condiciones monitorizadas, pero también permitiendo el análisis de escenarios ex-ante. Sin embargo, la preparación de los datos y la calibración de los modelos suelen ser tareas que requieren gran esfuerzo y experiencia, pudiéndose simplificarse dichas tareas con herramientas informáticas que automatizan parte del proceso. Se ha desarrollado una librería de R que agiliza la aplicación de un modelo de calidad del agua a cualquier cuenca fluvial. Incluye funciones para combinar fuentes de datos, evaluar las condiciones históricas, calibrar el modelo y evaluar las cargas y concentraciones anuales de nutrientes a lo largo de la red fluvial y en las salidas al mar, cuantificando las contribuciones de las fuentes difusas y puntuales a las cargas de nutrientes. Se presta especial atención a las funciones que permiten realizar la calibración y análisis de sensibilidad de los parámetros. Además, el paquete proporciona funciones para cartografiar las fuentes de nutrientes en una región, teniendo en cuenta las diferentes vías de acceso a las aguas y la estructura hidrológica de la red fluvial.

#medioambiente #ciencias #geográficas

CityShadeMapper: creando mapas de sombras de alta resolución espacio-temporal con datos abiertos de teledetección

Francisco Rodríguez Sánchez. *Departamento de Biología Vegetal y Ecología, Universidad de Sevilla, Sevilla, España*

Begoña Vega. *Innova-tsn*

María Neira. *Innova-tsn*

Ángela Díaz. *Innova-tsn*

En el contexto actual de cambio climático, la sombra es un recurso cada vez más necesario para garantizar la habitabilidad de las ciudades ante las elevadas temperaturas que se alcanzan gran parte del año. Al mismo tiempo, el sol es un recurso importante para muchos sectores, fundamentalmente en los meses de invierno.

Los autores han desarrollado un paquete de R para generar mapas de sombras de alta resolución espacio-temporal para cualquier municipio de España, utilizando datos abiertos de teledetección (LiDAR) ofrecidos por el Instituto Geográfico Nacional. El software permite conocer, para cada m2 de suelo y cualquier hora del año, el nivel de sombreado proporcionado por árboles, edificios, relieve e infraestructuras al nivel de la calle. Esta información puede ser tremendamente útil para mejorar la planificación urbana (p. ej. diseño de espacios públicos, elección de especies de plantas en calles y zonas verdes, ubicación de veladores, diagnóstico de zonas desprotegidas y creación de refugios térmicos, etc). El software también permite obtener “rutas de sombra” que maximizan la disponibilidad de sombra para desplazarse entre distintos puntos de la ciudad, aumentando el comfort térmico durante los meses de mayor rigor estival.

CityShadeMapper puede ser una herramienta fundamental para mejorar la planificación urbana y la adaptación de las ciudades y municipios al cambio climático.

#medioambiente #ciencias #geográficas

Interfaz de Monitorización - Shiny Dashboard integrada en Google Cloud

Jordi Segú Tell. *Gerencia de sistemas de información geográfica, Tragsatec, Madrid, Madrid, España*

Andrés Vinueza. *Departamento de Ciencia de Datos, LOGIKARESEARCH Cía. Ltda, Quito Ecuador*

Jorge Sosa. *Departamento de Matemática, Escuela Politécnica Nacional, Quito Ecuador*

La propuesta de comunicación se enmarca dentro de un proyecto de monitorización de las ayudas de la Política Agraria Común (PAC) mediante un análisis de los recintos declarados por los agricultores e imágenes de satélite, usando procesos automáticos y semiautomáticos. El resultado es la clasificación en tres categorías de tipo “semáforo” (verde, amarillo y rojo).

El semáforo se calcula con una clasificación predicha con Random Forest que determina la confianza entre el tipo de cultivo que declara un agricultor y el resultado de los indicadores espectrales extraídos mediante la teledetección en imágenes Sentinel. La interfaz de monitorización es una herramienta de trabajo interna que se utiliza después del Random Forest y sirve para:

- Visualizar los resultados obtenidos por el proceso de clasificación.
- Comparar los resultados de la clasificación automática con los resultados manuales que extraen los técnicos que van a campo y observan los cultivos de los agricultores. Así resulta posible extraer indicadores de falsos rojos y falsos verdes.
- Generación de hipótesis modificando los umbrales de la clasificación para aumentar o disminuir el equilibrio establecido de falsos rojos y falsos verdes.
- Validar la hipótesis generada.
- Crear resúmenes de las distintas fases del proyecto y enviarlos a través de correo electrónico.

La aplicación está creada íntegramente con el lenguaje de R a través de un dashboard de Shiny. El almacén de datos está alojado con el producto BigQuery de la nube de Google Cloud. Esta es una herramienta altamente escalable, que unida al dashboard de Shiny facilita el manejo de grandes cantidades de datos.

En la comunicación se explicarán las distintas funciones de la aplicación, la arquitectura de la misma y un ejemplo práctico de su uso.

#medioambiente #ciencias #geográficas

Adaptación de la metodología loci para la detección de anomalías en forma y magnitud para datos funcionales

Jorge Sosa Donoso. *Departamento de Matemáticas, Escuela Politécnica Nacional, Quito, Ecuador*

El presente trabajo desarrolla una metodología de detección de atípicos tanto en forma y magnitud para datos funcionales, adaptando el método Local Correlation Integral (LOCI), el cual, trabaja esencialmente con distancias y densidades para hallar las anomalías, por tanto, la adecuación de esta técnica se realiza mediante el cálculo de distancias en los espacios de Hilbert. Estudios de simulación son elaborados tomando en cuenta independencia o varios niveles de dependencia para las curvas simuladas, encontrando mejor desempeño cuando tenemos dependencia en los datos y si ésta es negativa, la detección de anomalías en magnitud es mejor, igualando los resultados cuando la dependencia es positiva. Finalmente, se aplica la metodología a un conjunto de datos sobre Humedad Promedio medidos por el Grupo de Energías Alternativas y Ambiente en Ecuador.

#medioambiente #ciencias #geográficas

Patrones de correlación espacial de la energía eólica en la península ibérica y sus implicaciones en planificación, optimización de portfolios y precios

Sergio Jiménez Sanjuán. *DNV, Energy Systems*

Francisco Angel Guerrero Vivas. *Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid, España*

Luis Martínez Uribe. *Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid, España*

La energía eólica y solar presentan una variabilidad temporal inherente debido a la naturaleza cambiante del viento. Sin embargo, existe una correlación entre la variabilidad del recurso entre diferentes localizaciones geográficas. Dependiendo del sentido e intensidad de estas correlaciones, las series agregadas de producción de diferentes plantas aumentarán o, en general, disminuirán su variabilidad relativa en mayor o menor grado. Entender los diferentes patrones de correlación entre diferentes localizaciones, que están relacionados con la naturaleza de los efectos climáticos locales, es de gran importancia para diferentes aplicaciones prácticas. Para un regulador es importante de cara a la planificación del crecimiento de la energía renovable. Para un propietario o inversor, para el diseño de portfolios que minimicen la variabilidad temporal de la producción y por tanto disminuyan los riesgos financieros. Para un agente representante, de cara al apantallamiento del desvío. También es relevante desde el punto de vista del precio efectivo de venta de la energía debido a la correlación existente entre el precio y el volumen total de energía renovable. A partir de datos de velocidad de viento e irradiancia procedente de modelos reanálisis climático ECMWF ERA-5, con una resolución de 0.25 x 0.25 grados longitud/latitud, y de Global Wind/Solar Atlas se han calculado las matrices de correlación espacial cruzada para cada posible combinación de celdas que cubren la península ibérica. Para entender bien los patrones de correlación calculados, es necesaria una simplificación regional. Para ello ha utilizado un algoritmo de clusterización jerárquico, utilizando las matrices de correlación como distancia. El resultado es una división automática del territorio en diferentes regiones que presentan patrones comunes de correlación en la evolución del recurso eólico y solar. Finalmente, mostraremos algunas aplicaciones de este análisis de correlación de cara a la optimización de portfolios, estimación de precios efectivos de la energía y analísis de ideoneidad para proyectos híbridos de energía solar y eólica.

#medioambiente #ciencias #geográficas

Manejo de datos climáticos multivariantes de alta resolución con el paquete “stars” para análisis espacio-temporales.

Daniel Romera Romera. *Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba, Córdoba, España*

Sandra Barragán. *Depto. Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

Un formato muy frecuente cuando utilizamos datos espaciales es el ráster (por medio de los paquetes para R “raster” y “terra”), el cual puede imaginarse como una cuadrícula de dos dimensiones (e.g., latitud y longitud) que contiene la información de una variable a lo largo de éstas. En análisis espaciotemporales, en los que se considera una tercera dimensión (i.e., tiempo), se suele trabajar con cuadrículas superpuestas, usando distintos formatos de cubo de datos (e.g., array, raster stack o raster brick). Pero ¿qué ocurre cuando nuestro estudio es más complejo (e.g., imágenes satelitales multibanda, o datos multivariantes)? Aunque estos formatos permiten manipular niveles crecientes de complejidad, pero su uso y manejo se vuelve más complicado y poco intuitivo. Además, es frecuente que los ecólogos nos topemos con limitaciones adicionales en nuestro trabajo con datos espacio-temporales multivariantes (e.g., capacidad de memoria y cómputo para el manejo de conjuntos de datos complejos y pesados). El paquete “stars” se presenta como una evolución del paquete “terra”, ofreciendo un formato de cubos de datos, diseñado para facilitar el manejo de datos espaciales-temporales multivariantes de gran tamaño y la integración con los paquetes de la familia “tidyverse”, aunque su reciente desarrollo hace que todavía sea poco conocido. En esta presentación se ilustra el uso del paquete “stars” para manejar datos mensuales de múltiples variables procedentes de simulaciones climáticas desde el Último Máximo Glacial (hace 22.000 años) hasta el año 2100 a alta resolución. Estos datos son de gran utilidad en estudios biogeográficos, pudiendo utilizarse en modelos de distribución, en estudios de series temporales de estas variables o de métricas derivadas (e.g., estabilidad climática en periodos concretos), e incluso como base para acciones de conservación de especies en el contexto del cambio climático. Además, se ilustrarán otras aplicaciones de interés en el ámbito de la ecología.

#medioambiente #ciencias #geográficas

Herramientas para el análisis geoespacial en R basado en áreas concéntricas para la determinación de emisión de aerosoles

Jesús Rojo. *Departamento de Farmacología, Farmacognosia y Botánica, Universidad Complutense de Madrid, Madrid, España*

Jaime Ballesteros de la Vega. *Unidad de Biometría, Sermes CRO, Madrid, España*

María Villar Navales. *Universidad Complutense de Madrid, Madrid, España*

Valeria Samira Samanamud Taus. *Universidad Complutense de Madrid, Madrid, España*

La calidad del aire viene determinada por los niveles de concentración de partículas y contaminantes inorgánicos o de tipo biológico presentes en las capas inferiores de la atmósfera. En este trabajo se presentan herramientas implementadas en R para el análisis de las fuentes potenciales de emisión de aerosoles, aplicados en diversos casos de estudio: desde el análisis del patrón de emisión polínica de especies anemófilas con respecto a la distribución y abundancia de sus fuentes; hasta los patrones globales de dispersión de material particulado procedente de áreas desérticas.

Este tipo de aproximación estadística permite realizar mapas continuos en formato Ráster de la concentración media de un determinado aerosol en el aire durante un periodo concreto de tiempo, lo cual se lleva a cabo en dos pasos principales: 1) se analiza el patrón de emisión de partículas mediante el método conocido como “Concentric Ring Method” [Oteros et al. 2015 J Environ Manage 155:212-218] implementado por primera vez en R, como una calibración del modelo basado en los datos de concentración de partículas procedentes de estaciones de las redes oficiales de monitorización de la calidad del aire; 2) el modelo generado se aplica a todos los píxeles en un área determinada mediante la caracterización de la distribución y la abundancia de las fuentes de emisión en anillos concéntricos.

Además de las herramientas de análisis, se muestran diversas soluciones para la visualización de datos espaciales basada en la comparativa de la abundancia y la distancia de las fuentes de emisión de aerosoles desde diferentes ubicaciones, como pueden ser las localizaciones de las estaciones de muestreo. Este estudio descriptivo de las fuentes de emisión puede resultar de gran interés como punto de partida en investigaciones científicas basadas en la calidad del aire respecto a partículas o contaminantes, tanto de origen biológico como inorgánico.

#medioambiente #ciencias #geográficas

Elaboración de un mapa de Códigos Postales de España con recursos libres: Cómo evitar pagar a Correos 6000€ por información de referencia

Francisco Goerlich. *Departamento de Análisis Económico e Instituto Valenciano de Investigaciones Económicas, Universidad de Valencia, Valencia, España*

Fuensanta Arnaldos-García. *Departamento de Métodos Cuantitativos para la Economía y la Empresa. Universidad de Murcia*

M. Teresa Díaz-Delfa. *Departamento de Métodos Cuantitativos para la Economía y la Empresa. Universidad de Murcia*

Un mapa (vectorial) oficial de Códigos Postales de España solo se puede conseguir si se le compra a Correos a un precio totalmente abusivo para tratarse de información de referencia generada con recursos públicos. ¡Más de 6000€ por una única descarga! Este trabajo suple esta carencia a partir de recursos libres. Fundamentalmente información sobre direcciones geo-codificadas publicadas por Cartociudad, y CNIG del IGN. El trabajo describe el algoritmo de elaboración de la capa vectorial de Códigos Postales, implementado en software libre, R. El algoritmo es sencillo y ajusta los Códigos Postales a los lindes municipales de la Base de Datos de Líneas Límite (BDLL) actual del IGN de forma exacta. El procedimiento es básicamente el siguiente. A partir de una capa vectorial puntual de direcciones con el Código Postal y de otra capa poligonal de contornos administrativos, las 3 operaciones básicas son: 1. Generar una teselación de Voronoi a partir de las direcciones. 2. Disolver la capa anterior por Código Postal. 3. Recortar el resultado anterior con la capa poligonal. El paso 1 convierte una geometría puntual en poligonal, manteniendo los atributos de cada punto. El paso 2 agrega los polígonos por Código Postal, y pierde el resto de los atributos. El paso 3 ajusta la geometría al contorno deseado. El proceso resultó laborioso porque la capa de direcciones geo-codificadas de Cartociduad no tiene un identificador directo de código municipal, y hubo que ajustar el proceso para solventar problemas de precisión geométrica en los lindes de los contornos administrativos y las direcciones de Cartociudad. La información generada se distribuye de acuerdo con la licencia de Cartociudad: Obra derivada de CartoCiudad 2006-2021 CC-BY 4.0. ¡Todo ello a un precio mucho más asequible que el solicitado por Correos, 0€! El trabajo completo, que incluye varios mapas de ejemplo, puede verse en https://www.uv.es/goerlich/Ivie/CodPost.html (https://www.uv.es/goerlich/Ivie/CodPost.html).

#bases #datos #geográficas

SDM-CropProj: una herramienta de modelación desarrollada en R para pronosticar la idoneidad ambiental de los cultivos y la producción de frutos

Salvador Arenas Castro. *Área de Ecología, Dept. de Botánica, Ecología y F. Vegetal, Universidad de Córdoba, Córdoba, España*

El cambio climático plantea desafíos importantes para la seguridad alimentaria mundial. Los cambios a largo plazo en la temperatura, la humedad, los patrones de lluvia y la frecuencia de los fenómenos meteorológicos extremos ya están afectando a las prácticas agrícolas, la producción de cultivos y la calidad nutricional de los cultivos alimentarios, lo cual tiene fuertes repercusiones socioeconómicas. Los esfuerzos globales para reducir las emisiones de gases de efecto invernadero y las medidas locales/regionales para mitigar y adaptarse a las condiciones climáticas cambiantes ya forman parte de las evaluaciones y actuaciones por parte de instituciones y autoridades globales que garanticen la seguridad de los alimentos, no sólo en relación con la salud y nutrición humana, sino también la salud animal/vegetal y el medio ambiente. Si bien el uso de escenarios futuros de cambio climático, a través de diferentes metodologías y protocolos de predicción, modelación o monitoreo como los Species Distribution Models (SDMs), ha sido una herramienta muy útil para anticiparse a los efectos del cambio climático en ecosistemas naturales, aún continúa siendo escasa en contextos agroecológicos. Anticipar estos cambios a través de la previsión de la superficie de cultivos ambientalmente adecuada, ayudaría a reducir o mitigar el impacto y adaptar las estrategias ecológicas y económicas para salvaguardar la seguridad alimentaria. En este sentido, presentamos aquí un protocolo asistido por modelos (en adelante, SDM-CropProj) íntegramente desarrollado en R, y que combina dos pasos principales de modelado que se implementarán en secuencia: 1) un proceso de calibración de múltiples técnicas y un enfoque de pronóstico por conjuntos para predecir el estado actual y la idoneidad ambiental futura de los cultivos objetivo; 2) un modelo lineal logarítmico univariante parsimonioso para relacionar la producción anual total promedio con el área adecuada actual basada en SDMs.

#medioambiente #ciencias #geográficas

Humboldt and rock varnish, a document-review in RPubs

José Jordán Soria. *Independent researcher*

Luis Martinez-Uribe. *Fundación Juan March*

Carlos Prieto. *Universidad de Salamanca*

David Barrios. *Universidad de Salamanca*

Cristina Calvo. *Universidad de Salamanca*

Friedrich Wilhelm Heinrich Alexander von Humboldt (1769-1859) was a prussian polymath with important scientific and cultural contributions, who inspiring naturalists like Charles Darwin with his exceptional holistic vision.

It is particularly interesting (and little known) that he was one of the first naturalist that described and studied mineral rock coatings in detail, being considerate the father of rock coating research.

On June 5th of 1799, both naturalists set sail from A Coruña (Spain) to Cumaná (Venezuela) beginning 5 years of expedition to America. During a year and a half explored the Amazonian tropical forest and the Orinoco River system. In the river basin, in waterfalls, he observed the presence of dark depositions on the rocks which he determined as a Mn-rich accretion (Humboldt, 1812).

This type of rock coating, nowadays called rock varnish, is composed by clay minerals (Si, Al) cemented on a bedrock enriched with variable quantities of Fe, Mn oxides and trace elements (eg. Ba, Ca, Ni) in its matrix but only it has been studied in detail in the last decades (Dorn, 2007 and Macholdt et al., 2017).

Nowdays is of great interest in astrobiology because (i) it has a high oxidative power, (ii) is gathers polyextreme conditions -pH, heavy metal accumulation, radiation-, (iii) it host microorganisms of biotechnological interest and (iv) it could contain biomarkers embebbed in its clay-matrix. In fact, interesting rock varnish-like coatings structures have been visualized in Mars and appreciable amounts of manganese in rocks have been detected by Chem-Cam on-board Curiosity (Lanza et al., 2014).

Therefore, the study of rock varnishes, which legacy was initiated by Humboldt, may be considered as an interesting terrestrial analogue to find extremal life and to be studied in much deeper for future years. An interactive document can be found at RPubs, in https://rpubs.com/Jose_Jordan/rock-varnish-review (https://rpubs.com/Jose_Jordan/rock-varnish-review).

#ciencias #vida

Salud y alimentación

Construcción de un flujo de extracción de valores clínicos recogidos de forma desestructurada empleando R: un ejemplo de uso con la fracción de eyección en pacientes con insuficiencia cardíaca

Tamy Goretty Trujillo Escobar. *Fundación Vasca de innovación e investigación sanitarias Bioef.*

Elena Rosa-Pérez. *Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España*

David Salgado. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

Introducción: La fracción de eyección (FE) es clave para decidir el tratamiento farmacológico de pacientes con insuficiencia cardiaca (IC). Ya que este dato no se registra de manera estructurada en Osakidetza- Servicio Vasco de Salud, se recurre a la revisión manual de informes de alta hospitalaria en los cuales los médicos acostumbran a recopilar el dato.

Objetivo: Desarrollar y validar un algoritmo de extracción basado en expresiones regulares con la herramienta R que permita la extracción automática de la FE de informes de alta hospitalaria de Osakidetza. Fundamentalmente, el algoritmo determina (i) la forma en la que se registra la FE, y (ii) en qué porcentaje de informes se encuentra.

Metodología: Creación del algoritmo: Empleando funciones propias y de diversos paquetes de R, se construye un flujo que a lo largo de iteraciones refina el algoritmo para (i) crear un diccionario con palabras clave, (ii) extraer y capturar líneas que contienen palabras del diccionario, y (iii) aplicar filtros y condiciones para depurar el resultado de la FE. Todos los criterios involucrados se definen con ayuda de personal médico. El algoritmo final arroja un valor expresado en número o cadena. Validación: sobre una muestra de 97 informes de alta de pacientes dados de alta tras un primer ingreso por IC entre 2014-2017 en Osakidetza, se compararon resultados del algoritmo versus la revisión manual de tres médicos especialistas.

Resultados: Sobre 88 informes con información relativa a la FE (en 11 no había mención a este valor) se obtuvo una sensibilidad 92,3%; especificidad 91,9%; VPP 92,3%; VPN 98,3%.

Conclusiones: El algoritmo presenta resultados favorables sobre una muestra de pacientes. Además, el proceso brinda información sobre la forma en que los médicos registran este tipo de valores clínicos. Dicho conocimiento es fundamental como paso previo a proyectos de procesamiento de lenguaje natural.

#ciencias #vida

Salud y alimentación: Jueves 10-11h (SaO2)

StudyDesign: Desarrollo de un entorno de simulación como herramienta de soporte al diseño de estudios y ensayos clínicos.

Alberto Sorribas. *Departamento de Ciencias Medicas Básicas, IRBLLEIDA, Universitat de Lleida, Lleida, España*

Gemma Pérez-López. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

José-Luis Zafra-Gómez. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

The design of a study is a critical phase that fixes the range of possible analyses that can be done and limits possible conclusions that come from the results. As such, the evaluating committees of projects, clinical trials, and publications pay special attention to the study design and often dismiss proposals that are not sufficiently reasoned and justified. Clinical and translational research groups often overcome this issue by consulting with a statistical service. In most studies different design alternatives can be considered, depending on factors such as access to patients, ease of follow-up, available budget, and the outcomes that are clinically relevant. In this crucial step, it is relatively common for the researchers to misunderstand the requirements or alternatives posed by statisticians. The project we propose will develop study simulation tools that allow an adequate discussion of the implications of a study design and facilitate interaction between clinical and statistical researchers. In doing so, we will contribute to the quality of the proposals and to enhance emerging groups with less experience in the design of studies, especially in areas such as primary medicine and nursing.

#ciencias #vida

Salud y alimentación: Jueves 10-11h (SaO3)

Automatización de sistema de vigilancia de incidencias COVID19 en centros educativos de la Comunidad Valenciana.

Carlos Abellán de Andrés. *Servicio de Evaluación y Estudios, Conselleria de Educación, Cultura y Deporte, Valencia, España*

Durante los cursos 2020-2021 y 2021-2022 el Servicio de Evaluación y Estudios de la Conselleria de Educación, Cultura y Deporte de la Generalitat Valenciana, implementó un sistema de vigilancia de la situación COVID19 en todos los centros educativos de la Comunidad Valenciana que impartieran educación infantil, primaria, secundaria, bachillerato y/o ciclos formativos.

El sistema de vigilancia constaba de cuatro procesos que se realizaban diariamente: 1. Obtención automática de los datos de incidencias COVID19 declaradas por los centros educativos mediante web scraping. 2. Tratamiento de dicha información y extracción de los resúmenes demandados por la Conselleria de Educación, Cultura y Deporte. 3. Generación de un dashboard mediante Rmarkdown en el que se mostraba la información más importante, así como gráficos interactivos utilizando leaflet y plotly. 4. Compartición de dicho dashboard de manera automática, primero mediante envío de correo electrónico (mailR) y posteriormente mediante la subida del fichero a un servidor local al que tenían acceso los destinatarios del resumen.

Además, en el proceso también se generaba de manera paralela: • Un pdf con la información mostrada en el dashboard para su uso en soporte papel si se consideraba necesario. • Un volcado de los datos a una planilla Excel demandada por el Ministerio de Educación y Formación Profesional. • Un Excel con la validación de los datos de origen mostrando posibles errores cometidos en la declaración de incidencias COVID por parte de los centros educativos.

El procedimiento se realizaba de manera automática a partir de la ejecución de un script base que, a su vez, iba ejecutando cada una de las tareas llamando a los diferentes archivos Rmarkdown.

#ciencias #sociales #humanidades

Estimación de la incidencia real de COVID-19 en España mediante un sistema de monitorización con R y Shiny

David Hervás Marín. *Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Valencia, España*

Begoña Vega. *Innova-tsn*

María Neira. *Innova-tsn*

Ángela Díaz. *Innova-tsn*

COVID-19 ha sido la primera pandemia que ha puesto de relieve la falta de preparación de las administraciones y sistemas sanitarios en todo el mundo frente a este tipo de fenómenos globalizados, con el consiguiente impacto económico y de vidas. En concreto, en España, hasta finales de julio de 2022 se han registrado más de 13.000.000 de casos positivos y más de 110.000 defunciones. Sin embargo, muchos de los casos reales no son registrados en el sistema o se registran con mucho retraso, con la consiguiente subestimación de la incidencia real. El objetivo de este trabajo es desarrollar un sistema monitorización y estimación de casos COVID-19 que será implementado en una aplicación web y que permitirá la cuantificación de la incidencia y el riesgo en tiempo real mediante la implementación de modelos estadísticos que proporcionarán información y predicciones sobre la evolución de variables de interés como el número de casos, la aparición y el número de brotes, las defunciones y los cambios de tendencia, entre otras, con el objetivo de facilitar la toma de decisiones informadas. Las estimaciones de estos modelos estadísticos servirán también para una mejor predicción de la demanda de recursos a los sistemas sanitarios y de la necesidad de implementación de restricciones y/o recomendaciones sanitarias a la ciudadanía. Este proyecto se enmarca dentro del ODS 3 “Salud y Bienestar”, concretamente con las líneas de actuación 3.8 “Lograr la cobertura sanitaria universal, en particular la protección contra los riesgos financieros, el acceso a servicios de salud esenciales de calidad y el acceso a medicamentos y vacunas seguros, eficaces, asequibles y de calidad para todos” y 3.d “Reforzar la capacidad de todos los países, en particular los países en desarrollo, en materia de alerta temprana, reducción de riesgos y gestión de los riesgos para la salud nacional y mundial”.

#epidemiología #salud #pública

Salud y alimentación: Jueves 10-11h (SaO5)

Estudio de poblaciones bacterianas en industria alimentaria mediante análisis estadístico multivariante

Adrián Alvarez Molina. *Departamento de Higiene y Tecnología de los Alimentos, Universidad de León, León, España*

Andrés Vinueza. *Departamento de Ciencia de Datos, LOGIKARESEARCH Cía. Ltda, Quito Ecuador*

Jorge Sosa. *Departamento de Matemática, Escuela Politécnica Nacional, Quito Ecuador*

Las nuevas tecnologías de secuenciación de ADN, junto al desarrollo de los métodos de computación, han posibilitado un enorme avance en la adquisición e interpretación de grandes volúmenes de datos.

Las materias primas y alimentos producidos por las industrias alimentarias entran en contacto con multitud de superficies antes de llegar al supermercado. El análisis de poblaciones microbianas en las superficies de industrias alimentarias es una herramienta que, aun estando todavía en fase de desarrollo, resulta prometedora para mejorar la seguridad alimentaria y evaluar la eficacia de procesos de limpieza.

En este trabajo se muestrearon superficies en contacto y no contacto con alimentos de 25 industrias alimentarias (3 mataderos, 10 cárnicas y 12 lácteas), asignándose las secuencias de ADN obtenidas a 1037 géneros bacterianos diferentes.

Objetivo del trabajo: aplicar técnicas estadísticas de análisis de datos multivariante para, en primer lugar, encontrar los géneros bacterianos característicos de distintas empresas, y en segundo lugar, cuantificar la influencia de las variables “tipo de industria” y “tipo de superficie” en la composición bacteriana.

Tras obtener la abundancia relativa de géneros bacterianos mediante procesos bioinformáticos, se evaluó programando en lenguaje R la similitud de poblaciones bacterianas mediante técnicas de clústering (clústering jerárquico, k-means), y de reducción dimensional: análisis de componentes principales (PCA), análisis de coordenadas principales (PCoA), escalado multidimensional no métrico (NMDS). Se determinó la influencia del tipo de industria y el tipo de superficie en la composición bacteriana mediante tests no paramétrico (adonis). Finalmente, con análisis discriminante lineal (LDA) se determinaron los géneros bacterianos característicos de cada tipo de industria y se corroboró con el análisis discriminante lineal con tamaño del efecto (LEFSe). Concluyéndose que el tipo de industria influyó más en la composición bacteriana que la superficie, siendo los géneros Lactococcus (lácteas), Psychrobacter (cárnicas y mataderos) o Pseudomonas (cárnicas) los más característicos.

#ciencias #vida

Salud y alimentación: Jueves 10-11h (SaO6)

Combinando shiny y MongoDB para facilitar el revisión sistemática de modelos de inactivación microbiana

Alberto Garre. *Departamento de Ingeniería Agronómica, Instituto de Biotecnología Vegetal, Universidad Politécnica de Cartagena (ETSIA), Cartagena, España*

Los modelos predictivos de inactivación microbiana son hoy en día una herramienta básica en la ciencia de alimentos, utilizándose en casos tan diversos como el análisis de riesgos, el diseño de proceso o en los estudios de vida útil. Hoy en día, los métodos para el desarrollo de estos modelos están bien establecidos. Por lo tanto, la revisión (sistemática) de la literatura es una parte esencial de prácticamente cualquier estudio. Sin embargo, la información en la literatura científica no se encuentra estructurada, por lo que extraerla es un proceso manual, trabajoso y susceptible de errores.

Por esta razón, hemos desarrollado D database. Esta aplicación web proporciona una interfaz a una base de datos de modelos de inactivación, así como funciones para el análisis de datos. La base de datos es el resultado de una revisión sistemática de la literatura y se ha compilado utilizando una arquitectura noSQL (MongoDB) hosteada en la nube (Atlas). D database se ha desarrollado en shiny y permite acceder a la base de datos realizando búsquedas de acuerdo a diferentes campos (microorganismo, producto…). Además, proporciona diversas funciones para el análisis de datos, tales como visualizaciones dinámicas, análisis estadístico o meta-análisis. La aplicación se ha desarrollado en código abierto y está disponible gratuitamente online (https://foodmicrowur.shinyapps.io/Ddatabase/ (https://foodmicrowur.shinyapps.io/Ddatabase/)).

La arquitectura de la aplicación se ha diseñado para facilitar su extensión y su mantenimiento. El uso de un sistema noSQL facilita la ampliación de la base de datos (p.ej. incluyendo otro tipo de tratamientos). Además, el código se ha escrito utilizando shinyModules, donde cada elemento (búsqueda de datos, visualización, ajuste de modelos…) es independiente. Esto facilita tanto el mantenimiento de los módulos actuales como el desarrollo de nuevas funciones.

#ciencias #vida

Comunicaciones Póster

Ciencias sociales y economía

Aplicaciones del aprendizaje automático automatizado en R

Mercedes Ovejero Bruna. *Unidad de Biometría, Sermes CRO, Madrid, España / Departamento de Metodología y Psicobiología, Universidad Complutense de Madrid, Madrid, España*

Elena Rosa-Pérez. *Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España*

David Salgado. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

El AutoML permite aplicar algoritmos de aprendizaje automático a problemas del mundo real, surgiendo como una solución basada en la inteligencia artificial para el desafío cada vez mayor que supone explorar modelos cada vez más complejos, ya sea por su alta dimensionalidad o bien por la calibración óptima de los hiperparámetros. Con el AutoML se abarca la elaboración completa de pipelines incluyendo desde el conjunto de datos sin procesar hasta el modelo final implementable. En R opciones de aplicación sencilla y eficiente del proceso del AutoML como el paquete H2O. En el presente trabajo, se postula que las estrategias de AutoML pueden representar un enfoque prometedor para mejorar la capacidad predictiva de los modelos tradicionalmente considerados en Ciencias Sociales y de la Salud. Para ello, se pone a prueba el análisis de una de las hipótesis centrales del modelo VIA de personalidad y fortalezas personales (Peterson y Seligman, 2004) según la cual estos rasgos predicen el bienestar. Esta hipótesis se puso a prueba hasta la fecha mediante procedimientos basados en los mínimos cuadrados ordinarios (OLS) como la regresión lineal múltiple. Para el presente trabajo, los instrumentos de evaluación de las variables objeto de estudio fueron el cuestionario VIA de fortalezas personales (evalúa 24 rasgos de personalidad) y el cuestionario de bienestar psicológico de Ryff (evalúa 6 componentes del bienestar psicológico). Ambos fueron aplicados a una muestra de 1274 estudiantes universitarios. Los resultados mostraron que: (1) el AutoML permitió construir el mejor modelo predictivo de forma eficiente; (2) se mejoró la capacidad predictiva de las variables de estudio en comparación con las aplicaciones tradicionales de modelos OLS; (3) es necesario que los datos tengan calidad y disponer de un conocimiento avanzado en modelización, así como en el área de investigación de interés para aplicar procedimiento de AutoML de forma adecuada.

#ciencias #vida

Digitalización de los centros educativos españoles a partir de la información de PISA 2018

M. Victoria Caballero Pintado. *Departamento de Métodos Cuantitativos para la Economía y la Empresa. Universidad de Murcia*

Gemma Pérez-López. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

José-Luis Zafra-Gómez. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

En los últimos años se puede observar una tendencia creciente en cuanto a la digitalización de todo tipo de procesos. Los centros educativos no son ajenos a ella y, en la gran mayoría, han desarrollado proyectos de digitalización que pretenden apoyar la denominada transformación digital educativa. Dado que el acceso a la digitalización por parte de los estudiantes no es igualitario, y depende, entre otros muchos factores, de su nivel socioeconómico, género, edad y familia, los centros educativos deben hacer posible el acceso a la competencia digital de todos los estudiantes por igual. No obstante, aún en un mismo país, el nivel de digitalización de los centros es diferente, ya que depende, entre otros aspectos, de la financiación del centro, su proyecto de digitalización, y la propia política educativa de la Comunidad Autónoma en la que se encuadra.

En este trabajo se emplea la información disponible en las encuestas de 2018 del programa PISA (programa de la OCDE de evaluación internacional de estudiantes) para clasificar a los centros educativos españoles participantes en perfiles con distinto grado de digitalización. La técnica empleada para la identificación es el análisis de perfiles latentes (LPA), que permite la agrupación de los centros atendiendo a sus características digitales a partir de la generación de una variable latente categórica que representa la pertenencia a un perfil que directamente no se puede observar. En el trabajo hemos podido clasificar a los centros educativos españoles que participan en PISA 2018 en un total de cuatro perfiles digitales, caracterizados por niveles diferentes tanto de disponibilidad digital en el centro como de orientación de su plan de digitalización. La asignación de los centros educativos a los distintos perfiles obtenidos permite la descripción posterior del proceso de digitalización llevada a cabo en cada una de la Comunidades Autónomas.

#ciencias #sociales #humanidades

Factores determinantes del precio de los alojamientos de Airbnb en Barcelona

Silvia Yepes Barbero. *Grado en Administración y Dirección de Empresas (Universidad de Murcia)*

La revolución tecnológica que ha acontecido en los últimos años ha hecho posible el desarrollo de nuevos modelos económicos, siendo uno de los más destacados el concepto de economía colaborativa. Consiste en el intercambio peer-to-peer de cualquier bien o servicio mediante diversas plataformas online. De los muchos ejemplos disponibles, nos hemos centrado en el análisis de Airbnb, una popular plataforma online de alojamientos turísticos, y en la determinación de los factores que afectan al precio de dichos alojamientos. Hemos analizado los datos obtenidos de la web para la ciudad de Barcelona, en junio de 2021, teniendo en cuenta dos tipos de alojamientos: la vivienda completa, que suele ser alquilada por consumidores que viajan en grupo, y las habitaciones privadas, populares entre turistas con bajo presupuesto que suelen viajar individualmente o en grupos más pequeños.

El estudio de los factores que afectan al precio se ha realizado empleando modelos de precios hedónicos. Se ha dividido en cuatro modelos diferentes para cada uno de los dos tipos de alojamiento dependiendo de las características de las variables que los componen: características estructurales del alojamiento, características de los anfitriones, prestaciones disponibles y reseñas de los huéspedes.

Finalmente, se ha tenido en cuenta el componente geográfico. El análisis de la localización de los alojamientos nos ha indicado que muchos de ellos están concentrados en los barrios Ciutat Vella y Eixample, las zonas más turísticas de la ciudad. Los alojamientos que se ofrecen en estas ubicaciones tienen los precios más altos, efecto que también se ha incluido en el modelo. Los resultados muestran que son las variables relacionadas con el tamaño del alojamiento y los servicios extra ofrecidos por el anfitrión los que tienen un efecto sobre el precio superior al resto.

#ciencias #económicas

Redes interactivas para el análisis datos con el paquete netCoin

Modesto Escobar. *Universidad de Salamanca*

Begoña Vega. *Innova-tsn*

María Neira. *Innova-tsn*

Ángela Díaz. *Innova-tsn*

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. El paquete netCoin permite la generación de gráficos interactivos que proporcionan al análisis unas posibilidades exploratorias de las que carecen muchas otras técnicas. 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.

#ciencias #sociales #humanidades

Docencia y programación R

Programación matemática en R

José Antonio Martín Fernández. *Depto. Informática, Matemática Aplicada y Estadística, Universitat de Girona, Girona*

Elena Rosa-Pérez. *Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España*

David Salgado. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

La limitación de todo tipo de recursos hace que hoy en día sea más importante que nunca la optimización de su consumo. La programación matemática ofrece herramientas para la optimización de sistemas organizativos definidos mediante variables deterministas. Este tipo de sistemas organizativos son muy habituales en el campo de las ingenierías y la informática. Las librerías para optimización en R junto con la posibilidad de la propia programación de funciones hacen de R y RStudio un entorno idóneo para introducir estas técnicas en los cursos avanzados de las ingenierías. El paquete “R Optimization Infrastructure” (ROI: http://roi.r-forge.r-project.org (http://roi.r-forge.r-project.org)) proporciona una estructura dónde se integran 19 librerías de optimización cubriendo desde la programación lineal (paquete “ROI.plugin.lpsolve”) hasta la programación no-lineal general (paquete “ROI.plugin.alabama”). Cada librería del lenguaje R dedicada a la resolución de problemas de optimización tiene su propia estructura de funciones, que tienen diferentes parámetros de entrada y distintos campos en los registros de las salidas de resultados. Sin embargo, los problemas de optimización (PL, PE, PQ, PNL) tienen unos elementos básicos comunes en su modelo: la función objetivo. y las restricciones. También comparten unos resultados básicos en su resolución: valor de la función objetivo, valor de las variables de decisión y de las holguras. El paquete ROI está diseñado para integrar en un único paquete todas las librerías de funciones para resolver problemas de optimización en un único entorno donde las llamadas a funciones, los parámetros de entrada y los registros de salida de resultados estén unificados. Tenemos a nuestra disposición una ventaja adicional, el paquete ROI es extensible con nuestras propias funciones o paquetes de optimización. En este trabajo se presenta cómo introducir las técnicas de la programación matemática en el entorno RStudio mediante la confección de informes en R-Markdown.

#matemáticas

Estadística y análisis de datos

Ejemplo de investigación operativa con R. Un problema de localización

Jose Luis Cañadas Reche. *Orange Spain, Madrid, España*

Elena Rosa-Pérez. *Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España*

David Salgado. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

Gracias a la versalitidad de R , se puede utilizar algoritmos y software desarrollados fuera de este. En esta comunicación veremos como utilizar “solvers” de COIN-OR a través de librerias como ROI o ompr. Para eso vamos a ver como resolver un problema típico de investigación operativa como el de asignar usuarios a localizaciones de forma óptima y sujeto a varias restricciones.

#matemáticas

El concepto de potencia estadística en análisis de la varianza

Pedro Sandoval. *Departamento de Ciencias Médicas Básicas, IRBLLIDA, Universitat de Lleida, Lleida, Espanya*

Gemma Pérez-López. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

José-Luis Zafra-Gómez. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

El concepto de potencia es difícil de entender para muchos usuarios. En particular, su cálculo y la adecuada comprensión en la técnica del análisis de la varianza no es fácil. Mediante simulación podemos profundizar en su significado y en las implicaciones prácticas. El diseño de una herramienta basada en shiny facilita la tarea y proporciona una aproximación intuitiva al problema.

#ciencias #vida

LAMPR: Linux, Apache, MySQL, PHP y R. Arquitectura Web para el desarrollo de plataformas Web de análisis.

Saúl Pastor. *Servicio de Bioinformática, Nucleus, Universidad de Salamanca.*

El desarrollo de aplicaciones Web de análisis de datos permite la ejecución de métodos analíticos a usuarios sin conocimientos previos en ciencia de datos. En el entorno R, el desarrollo de aplicaciones con Shiny permite la ejecución de programas R dentro de una interfaz Web, sin embargo, para el desarrollo de aplicaciones avanzadas es complicado adaptar Shiny a los estándares de desarrollo Web actual. En este resumen proponemos una arquitectura de desarrollo que hemos denominado LAMPR (Linux, Apache, MySQL, PHP, R) adecuada para desarrolladores Web que trabajen en plataformas Web de análisis de datos. El empleo del servidor Web Apache aporta seguridad y robustez a la aplicación y permite hospedar múltiples proyectos por el puerto 80. El sistema operativo Linux aporta ventajas a nivel de seguridad y permite la ejecución de comandos y programas para el tratamiento de datos propios de Linux. Mediante PHP y MySQL se consigue desarrollar páginas Web dinámicas y almacenar datos preprocesados y resultados en el tiempo sin necesidad de realizar nuevas ejecuciones. También facilitan el empleo de sesiones, cuentas de usuario y la realización de gestores de procesos. Finalmente, la conexión de PHP con R consigue la ejecución de procesos de análisis de datos y el manejo eficiente de una gran cantidad de información en memoria. Esta arquitectura ha sido empleada en el desarrollo de servidores como https://mutationmining.usal.es/ (https://mutationmining.usal.es/) , https://ranaseq.eu/ (https://ranaseq.eu/) o https://singlecanalyzer.eu/ (https://singlecanalyzer.eu/).

#ciencias #informáticas #ingeniería

Medioambiente y ciencias geográficas

Identificación de elementos similares a retrotransposones en el genoma de un patógeno fúngico utilizando KaryoploteR

María Victoria Aguilar Pontes. *Departamento de Genética, Universidad de Córdoba, Campus Universitario de Rabanales, Edif C5, E-14071 Córdoba, España*

Elena Rosa-Pérez. *Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España*

David Salgado. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

Los retrotransposones (REs) son elementos genéticos capaces de amplificarse a sí mismos a través de un intermediario de ARN y contribuyen a la plasticidad de los genomas, influyendo en la evolución y adaptación de las especies. Se han identificado REs en casi todos los organismos eucariotas. En general, los REs están silenciados, pero en situaciones de estrés pueden ser activados. Están compuestos por una secuencia larga de entre cientos y varios miles de pares de bases transcritas como un único ARNm que codifica múltiples enzimas, tales como la transcriptasa inversa, utilizadas durante su propagación. La identificación de REs en genomas es un desafío debido al alto número de copias de secuencias casi idénticas y las estructuras complejas producidas por nuevas inserciones de REs en secuencias de REs existentes. Usando el paquete KaryoploteR visualizamos estructuras similares a REs en el genoma de Fusarium oxysporum f. sp. lycopersici (Fol), un hongo patógeno responsable de la marchitez vascular en más de 150 cultivos vegetales que puede infectar también huéspedes animales, incluso humanos. En un análisis de RNA-seq de Fol durante la infección de plantas se observó un aumento claro en las lecturas que mapeaban fuera de los transcritos anotados en el genoma. Utilizando el paquete KaryoploteR se seleccionaron las lecturas fúngicas que sólo mapean a nivel de genoma pero no de transcriptoma y se visualizó la cobertura del genoma completo. De esta forma identificamos varias regiones que cubren hasta 500.000 bp con alta cobertura y que corresponderían con estructuras en tándem similares a REs. Un análisis con Blast de dichas regiones mostró homología con proteínas involucradas en la propagación de REs. En resumen, el uso de KaryoploteR permitió visualizar de una manera fácil y rápida la cobertura a nivel de genoma completo e identificar regiones activas que codifican supuestos REs.

#ciencias #vida

Estimación de parámetros forestales mediante LiDAR Terrestre: Aplicaciones en R

Antonio Jesús Ariza Salamanca. *Departamento de Ingeniería Forestal, Universidad de Córdoba, Córdoba, España*

Gemma Pérez-López. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

José-Luis Zafra-Gómez. *Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España*

El láser escáner terrestre (TLS; Terrestrial Laser Scanner) permite la obtención de nubes de puntos tridimensionales (representación discreta) muy próximas a superficies continuas, basándose en el tiempo que el haz láser tarda en interceptar un objeto y volver al instrumento. Este dispositivo ha sido empleado en múltiples aplicaciones, destacando su gran potencial en el ámbito forestal debido a la medición rápida y no destructiva que hace de los ecosistemas forestales. El TLS posee la capacidad de registrar con un alto nivel de detalle la estructura vertical y horizontal de una parcela forestal, lo cual ha provocado un enorme interés en la implementación de este dispositivo para la obtención de variables forestales. No obstante, dado que los datos generados por el TLS pueden comprender millones de puntos para una sola parcela de medición, se necesitan métodos sofisticados para procesarlos de forma automática. En este sentido, numerosos algoritmos capaces de extraer parámetros forestales a partir de datos procedentes de TLS han sido desarrollados en softwares de acceso libre como es el caso del lenguaje de programación R. El repentino aumento del número de actualizaciones de paquetes de R desarrollados con el objetivo de automatizar el procesamiento de nubes de puntos de TLS para la estimación de variables de interés forestal, ejemplifica la creciente importancia del lenguaje R en el sector forestal y el fortalecimiento de la comunidad. Los objetivos de esta comunicación son: (1) presentar las aplicaciones actuales del TLS para la estimación de parámetros forestales y las aplicaciones resultantes en términos de gestión o ecología forestal; (2) describir algunos de los principales paquetes de R que han sido desarrollados para la estimación de estos parámetros y; (3) enumerar algunas de las direcciones futuras para fortalecer la comunidad. Se presentan los paquetes: FORTLS, AMAPVox y rTLS, junto con sus principales aplicaciones.

#medioambiente #ciencias #geográficas

Preparación y comparación de escenarios climáticos actuales y futuros: un flujo de trabajo ilustrativo en R

Claudio Açaí Bracho Estévanez. *Depto. de Biología, IVAGRO, PaisajeLAB – Febimed Group, University of Cádiz, Campus Río San Pedro, Puerto Real. Cádiz, Spain*

En un contexto de calentamiento global antropogénico, la predicción de escenarios climáticos futuros es fundamental tanto para investigaciones teóricas como aplicadas. Los modelos climáticos desarrollados en la actualidad y basados en la simulación de flujos de materia y energía entre compartimentos geofísicos (atmósfera, océanos y tierra), proporcionan proyecciones climáticas con diverso grado de incertidumbre. Existe una gran variedad de modelos climáticos desarrollados por distintas instituciones que, varían según la metodología, el escenario de emisión o la resolución temporal de una predicción determinada. Por tanto, operar con predicciones climáticas actuales y/o futuras en un área de estudio definida requiere algunos pasos de preparación y análisis previo. La mayoría de esos pasos pueden llevarse a cabo en R mediante el uso diversos paquetes (p.ej. raster) para finalmente comparar y operar con variables bioclimáticas (generalmente disponibles como datos raster). En este póster presentaremos un flujo de trabajo ilustrativo para preparar y comparar proyecciones climáticas actuales y futuras en un marco geográficamente explícito. Se utilizaron tres modelos (IPSL, MPI y UKESM), tres escenarios de emisión (ssp126, ssp370 y ssp585) y dos resoluciones temporales (2030 y 2090) para construir proyecciones futuras a nivel europeo. Además, proporcionaremos algunas comparaciones geográficas entre escenarios futuros y actuales para visualizar cómo el cambio climático tiene efectos distintos a través de la dimensión espacial de nuestra área de estudio.

#medioambiente #ciencias #geográficas

KTU2: partitioning KTU clustering algorithm for highly diverse microbiome data

Miguel Camacho Sánchez. *Instituto Andaluz de Investigación y Formación Agraria, Pesquera, Alimentaria y de la Producción Ecológica*

Begoña Vega. *Innova-tsn*

María Neira. *Innova-tsn*

Ángela Díaz. *Innova-tsn*

El metabarcoding se ha popularizado para caracterizar la diversidad biológica de muestras ambientales y clínicas. Consiste en la secuenciación de una región concreta del ADN total de la muestra, que hace de “código de barras” a través de su comparación con bases de datos de secuencias conocidas de taxones. El flujo bioinformático elegido para analizar estos datos masivos tiene un gran impacto en las inferencias biológicas. La mayoría confluyen en un paso clave de determinación de variantes singulares (ASVs): secuencias únicas que representan a un taxón concreto. Liu et al. (2021) (10.1111/2041-210X.13758) desarrollaron recientemente KTU, un paquete de R que permite agregar los ASVs basándose en K-meros (K-mer Taxonomic Units, KTU), consiguiendo exitosamente una reducción en los ceros en la matriz de abundancia de taxones y mejorando consecuentemente la interpretación biológica de los resultados de metabarcoding. No obstante, esta primera versión de KTU tiene dos limitaciones significativas que dificultan su uso práctico: 1) explota intensamente la memoria RAM, volviéndolo inoperativo en equipos pequeños, y 2) las funciones internas solamente admiten como input los resultados generados en QIIME. La versión KTU2, en desarrollo (github.com/poyuliu/KTU2), incorpora la nueva función ktusp que hace una partición previa de los datos en grupos, usando un árbol filogenético como guía, antes de generar los KTUs. Ello permite procesar los datos de manera seriada, controlando el uso de la RAM. Además, se han creado nuevas funciones que permiten correr el algoritmo principal de clustering a partir de matrices de datos de ASV generadas en R por los populares paquetes usados en metabarcoding dada2 y phyloseq. Este último punto permite un procesado completo de las muestras sin salir del entorno de R.

#ciencias #vida

Cálculo del Índice de Competencia Forestal para parcelas del Inventario Forestal Nacional

Aurelio Diaz Herraiz. *Departamento de Botánica, Ecología y Fisiología Vegetal. Universidad de Córdoba, Córdoba, España. Instituto Federal de Ciência e Tecnologia do Amazonas, Campus Humaitá, Amazonas, Brasil*

Andrés Vinueza. *Departamento de Ciencia de Datos, LOGIKARESEARCH Cía. Ltda, Quito Ecuador*

Jorge Sosa. *Departamento de Matemática, Escuela Politécnica Nacional, Quito Ecuador*

Actualmente, muchos estudios analizan el crecimiento de los árboles en función de variables abióticas como el clima, el suelo o la topografía descuidando el efecto que variables bióticas como la densidad forestal o el tamaño de los individuos pueden ejercer sobre la biomasa forestal. La disponibilidad de recursos como la luz, el agua y los nutrientes depende de la densidad de individuos compitiendo por ellos. Sin embargo, la densidad no siempre describe el grado de interacción entre los árboles. La combinación de factores como la distancia entre individuos vecinos, su diámetro basal y el área de influencia de cada árbol pueden formular un índice que estime la competencia (IC) el cual afecte al crecimiento del individuo. El IC del árbol i se define como la sumatoria de los efectos generados por todos los árboles que rodean al árbol i en un radio de 7 metros. El efecto individual de los j arboles se define como el cociente de dividir el diámetro del árbol j entre el diámetro del árbol i dividido por la distancia entre ambos. Este trabajo aborda el cálculo del IC no como una operación condicional (que excluyan secuencialmente los árboles a más de 7 metros de cada árbol antes de la operación sino partiendo de matrices que calculan todos los efectos independientes sobre cada uno de los árboles simultáneamente antes de filtrar los árboles a más de 7 metros. Esto permite mayor rapidez y escalabilidad del cálculo posibilitando su fácil replicación en un mayor número de árboles sin consumir un alto poder computacional. Esta metodología en R sigue la tendencia gramática actual que busca excluir bucles en lenguajes de programación (como Python y Matlab) ayudando a buscar soluciones alternativas en la algebra lineal.

#medioambiente #ciencias #geográficas

Genotyping-By-Sequencing (GBS) Analysis of traditional Spanish melon landraces

Alejandro Flores León. *Instituto de Conservación y Mejora de la Agrodiversidad Valenciana (COMAV), Universitat Politècnica de València, Valencia, Valencia, España*

Melon (Cucumis melo L.) is an important and diverse crop of the Cucurbitaceae family, with Spain being secondary centre of diversification for this crop. Most recently, its traditional highly diverse landraces are being replaced by F1-Hybrids belonging to the Piel de Sapo and Amarillo subgroups. To understand the biodiversity of melons in Spain, traditional landraces were studied belonging to Ibericus and Flexuosus groups (research projects PID2020-116055RB C21, PROMETEO 2017/078, PROMETEO/2021/072). A total of 38 traditional landraces (2 Exotic Indian, 1 Chate, 4 Flexuosus and 31 Ibericus) were analysed employing GBS approach, obtaining SNPs to study their genetic diversity. A total of 25529 quality SNPs (maf 0.05, minimum 10 counts, max-missing 4) were obtained. Population Structure was analysed, resulting in K=2 being the best result, with K=3 being the second-best result. The subpopulations observed was one formed by the Spanish Ibericus sweet melons, and the other by the Spanish Flexuosus melons, with Chate and Exotic landraces being between the two (K=2). The Linkage disequilibrium (LD) decay was studied, varied significantly between Ibericus and Flexuosus and Chate group. The phylogeny was performed employing the exotic Kachri landrace as a root. The phylogenetic results showed that Flexuosus and Ibericus formed their own separate clades. Among Ibericus, no clear grouping by their subgroups was seen, although Piel de Sapo and some Tendrales and Amarillos did present very little distance between them. This study showed the genetic diversity between different traditional melon Spanish landraces, both sweet and non-sweet.

#genetica #vegetal

Simulaciones inversas en Hydrus-1D a través de R para caracterizar el movimiento de agua en el suelo

Ismael Lare David. *Departamento de Ciencias Agrarias y del Medio Natural, Universidad de Zaragoza, Zaragoza, España*

Francisco Angel Guerrero Vivas. *Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid, España*

Luis Martínez Uribe. *Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid, España*

Los procesos de carga y descarga de agua a través de o desde el suelo son determinantes para conocer la disponibilidad de agua para la vegetación, así como el almacenamiento de agua en el perfil del suelo o en los acuíferos. Para tener un conocimiento detallado de los mismos se recurre tanto al seguimiento como al modelado de los flujos y variables de estado fundamentales del proceso, como, por ejemplo, la humedad del suelo, la tensión o la percolación profunda. HYDRUS-1D (Simunek et al., 1998) es un software que simula el flujo y transporte unidimensional de agua, calor y solutos teniendo en cuenta las ecuaciones de difusión, dispersión y transporte. Para resolverlas tiene en cuenta unas condiciones de contorno, superior a partir de datos meteorológicos (precipitación, evaporación y transpiración) e inferior (drenaje, flujo constante, etc.), y unos parámetros que regulan la retención y transmisión de agua y solutos. Estos parámetros pueden obtenerse bien mediante el muestreo y análisis en laboratorio o bien mediante simulación inversa. Este último procedimiento consiste en la repetición sistemática de simulaciones del modelo con conjuntos de parámetros diferentes obtenidos del muestreo aleatorio de los mismos de entre sus posibles valores. Las simulaciones son comparadas con las observaciones para, de este modo, encontrar un intervalo de parámetros concreto que minimice el error de la estimación. En este trabajo el proceso iterativo, de minimización del error de la estimación del modelo Hydrus-1D mediante el método de mínimos cuadrados, se ha implementado en un código en R. Para ello, empleando el paquete hydrusR (v 0.3.0; Subodh, 2020) como punto de partida, el cual realiza modelos directos ejecutando Hydrus 1-D, se han desarrollado nuevas funciones que permitan ejecutar simulaciones inversas y repetir el proceso para distintos periodos de simulación que den relevancia estadística al proceso. El código desarrollado se estructura en: i) creación de una carpeta donde irán destinados todos los archivos generados por la simulación inversa del modelo; ii) generación de archivos con datos de entrada al modelo inverso (meteorología y medidas de humedad del suelo) para cada periodo de simulación; iii) ejecución del modelo inverso; iv) lectura de los resultados de la simulación y comparación de los parámetros ajustados con los obtenidos en el resto de simulaciones inversas; v) representación gráfica de los modelos ajustados y datos observados. Por lo tanto, el código desarrollado es capaz de automatizar la calibración del modelo, proporcionando intervalos de incertidumbre para los parámetros obtenidos que puedan ser empleados posteriormente para analizar escenarios futuros o eventos registrados y, de este modo, conocer mejor el sistema hidrológico en estudio y realizar una gestión eficiente del mismo. Referencias. Simunek, J., Huang, K., van Genuchten, M.T.Th., 1998. Hydrus - code for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media.The HYDRUS code for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media. Versio, Research R. ed. U.S. Salinity Laboratory, USDA-ARS, Riverside, California. Subodh, A. 2020. hydrusR: Utiliy package to run HYDRUS-1D and analyse results. R package version 0.3.0.

#ciencias #vida

Integración de LIDAR y Sentinel-2 para el inventario forestal dinámico

Santiago Martin Malcon. *Agresta S. Coop.*

Sandra Barragán. *Depto. Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

El objetivo principal del proyecto es el desarrollo de un prototipo de plataforma de inventario forestal dinámico de superficies forestales mediante el uso de fuentes de datos capturados mediante sensores remotos (LiDAR y Sentinel-2), empleando técnicas de aprendizaje automático, estadística avanzada y análisis espacial, para estimar un amplio espectro de variables biofísicas con las que caracterizar de manera cualitativa y cuantitativa las masas forestales de una forma dinámica –mediante la actualización periódica a través de la incorporación de los cambios ocurridos en las superficies objeto de inventario–. LiDARBosc se sustenta en el uso de datos de libre disposición y software Open Source (FUSION, R y Shiny fundamentalmente). Concretamente, utilizará datos LiDAR del PNOA (Plan Nacional de Ortofotografía Aérea) del Instituto Geográfico Nacional), imágenes del satélite Sentinel-2 del programa Copernicus de la Unión Europea, y datos del Inventario Forestal Nacional (Ministerio de Agricultura, Pesca y Alimentación).

#medioambiente #ciencias #geográficas

Salud y alimentación

Análisis espacial de las paradas cardiacas extrahospitalarias atendidas en España

Patricia Fernández del Valle. *Comité de Ética de la Investigación (CEIm) de Cáceres, Fundesalud, Complejo Hospitalario Universitario de Cáceres, Cáceres, España*

Elena Rosa-Pérez. *Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España*

David Salgado. *Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España*

Introducción_La Parada Cardiorespiratoria(PCR) es un grave problema de salud pública.En Europa se considera una de las principales causas de muerte.Existe gran diferencia tanto en la incidencia como en la supervivencia de PCR entre países.En España también existe variabilidad.Los objetivos de este estudio son conocer los factores asociados a la supervivencia a PCRExtrahospitalaria(PCRE) y examinar su variabilidad en España. Métodos_Se trata de estudio descriptivo basado en Registro Español de Parada Cardiaca de España(OSHCAR).Se incluyeron casos correspondientes al primer periodo OHSCAR con etiología identificada como médica.Se recogieron variables relativas al paciente,evento,asistencia previa,asistencia realizada por el Equipo Emergencias(EE),y al resultado final en términos de supervivencia. Se realizó un análisis descriptivo de las variables por comunidades autónomas(CCAA).Se examinaron los factores asociados a la supervivencia a través de regresión logística.Se elaboró una representación espacial de las variables en cada CCAA utilizando mapas de coropletas.Se exploró existencia de dependencia espacial utilizando estadístico I-Moran. Resultados_Se incluyeron 7785 pacientes.La edad mediana fue 67.0 años[IQR34.5-77.9].El 72.1% fueron hombres y 60.0% ocurrieron en el domicilio. El 75.6% fueron presenciadas y en el 56.8% se realizaron maniobras RCP antes de llegar EE.El 24.1% presentaron ritmo inicial desfibrilable y en el 25.7% el EE llegó antes de 8” de la llamada.El 30.2% de los pacientes fueron trasladados con ROSC al hospital(rango:17.3-50.0).Recibieron el alta hospitalaria el 11.7% (38.5% de los que llegaron con ROSC al hospital).El 10.2% con buen estado neurológico,CPC1-2(33.6% de los que llegaron con ROSC al hospital). Las variables asociadas a supervivencia fueron:edad,sexo(masculino),motivo de llamada(sospechaPCR),PCRE presenciada,PCR en el domicilio,ritmo inicial desfibrilable,intervalo llamada-llegada(=8”) y CCAA. La autocorrelación espacial proporcionó I-Moran=0.3324(p=0.0129) para ROSC al hospital. Conclusiones_A pesar de existir una estructura similar en servicios de emergencias extrahospitalarios españoles,existe una importante variabilidad de la PCRE.La representación gráfica a través de mapas, permite visualizar de forma sencilla las diferencias existentes.Disponer datos a un nivel geográfico menor permitiría un análisis más exahustivo.

#ciencias #vida

Listado de autores

Nombre	Afiliacion	Comunicaciones
Adrián Alvarez Molina	Departamento de Higiene y Tecnología de los Alimentos, Universidad de León, León, España	SaO5
Aguilar-Pontes, M.V.	Departamento de Genética, Universidad de Córdoba, Campus Universitario de Rabanales, Edif C5, E-14071 Córdoba, España	MeP1
Aida Calviño Martinez	Departamento de Estadística y Ciencia de los datos, Facultad de Estudios Estadísticos, Universidad Complutense, Madrid, España	DoO10
Albert Sorribas	Departamento de Ciencias Médicas Básicas, IRBLLIDA, Universitat de Lleida, Lleida, Espanya	EsP2
Alberto Aloe	Joint Reseach Centre, Ispra, Italy	MeO3
Alberto Garre	Departamento de Ingeniería Agronómica, Instituto de Biotecnología Vegetal, Universidad Politécnica de Cartagena (ETSIA), Cartagena, España	SaO6
Alberto Morillo-Alujas	Tests & Trials, S.L.U., grupo Tentamus	EsO12
Alberto Sorribas	Departamento de Ciencias Medicas Básicas, IRBLLEIDA, Universitat de Lleida, Lleida, España	SaO2
Alejandro Flores-León	Instituto de Conservación y Mejora de la Agrodiversidad Valenciana (COMAV), Universitat Politècnica de València, Valencia, Valencia, España	MeP6
Ana C. Cebrián	Departamento de Métodos Estadísticos, Universidad de Zaragoza, Zaragoza, España	MeO2
Ana Martina Greco	Estudis de Dret i Ciència Política, Universitat Oberta de Catalunya, Barcelona, España	EsO6
Ana Pérez-de-Castro	Instituto de Conservación y Mejora de la Agrodiversidad Valenciana (COMAV), Universitat Politècnica de València, Valencia, Valencia, España	MeP6
Andrea Martín Rodríguez	Gerencia de sistemas de información geográfica, Tragsatec, Madrid, Madrid, España	MeO5
Andrés Vinueza	Departamento de Ciencia de Datos, LOGIKARESEARCH Cía. Ltda, Quito Ecuador	CiO6
Angel Udias Moinelo	Dpto. Ciencias de la Computación, Arquitectura de Computadores, Lenguajes y Sistemas Informáticos y Estadística e Investigación Operativa, Universidad Rey Juan Carlos, Mostoles, Madrid, Españ	MeO3
Ángela Díaz	Innova-tsn	CiO3

Nombre	Afiliacion	Comunicaciones
Antonio Jesús Ariza Salamanca	Departamento de Ingeniería Forestal, Universidad de Córdoba, Córdoba, España	MeP2
Antonio Picornell	Departamento de Botánica y Fisiología Vegetal, Universidad de Málaga, Málaga, España	MeO10
Aurelio Diaz Herraiz	Departamento de Botánica, Ecología y Fisiología Vegetal. Universidad de Córdoba, Córdoba, España. Instituto Federal de Ciência e Tecnologia do Amazonas, Campus Humaitá, Amazonas, Brasil	MeP5
Aurora González-Vidal	Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, España e ITI-CERTH, Salónica, Grecia.	EsO13
Avelino Alvarez Ordóñez	Departamento de Higiene y Tecnología de los Alimentos, Universidad de León, León, España	SaO5
Bart Jan van Os	Leiden University	EsO2
Begoña Vega	Innova-tsn	CiO3
Belen Picó	Instituto de Conservación y Mejora de la Agrodiversidad Valenciana (COMAV), Universitat Politècnica de València, Valencia, Valencia, España	MeP6
Bojan Mihaljevic	Universidad Politécnica de Madrid	EsO17
Bruna Grizzetti	Joint Reseach Centre, Ispra, Italy	MeO3
Carlos Abellán de Andrés	Servicio de Evaluación y Estudios, Conselleria de Educación, Cultura y Deporte, Valencia, España	SaO3
Carlos de la Calle Arroyo	Instituto de Ciencia de los Datos e Inteligencia Artificial, Universidad de Navarra, Pamplona, España	EsO1
Carlos Ortega Fernández	QUALITYEXCELLENCE SL, MADRID, ESPAÑA	NAT3
Carlos Prieto	Universidad de Salamanca	CiP4, EsP3
Carlos Prieto	Servicio de Bioinformática, Nucleus, Universidad de Salamanca.	CiP4, EsP3
Carlos Prieto Sánchez	Servicio de Bioinformatica, Nucleus, Universidad de Salamanca.	EsO16
Catalina García García	Universidad de Granada, Granada, España	EsO9
Cesar Alfaro	URJC, Mostoles, Madrid	MeO3
Clara Peréz Moro	Instituto de Conservación y Mejora de la Agrodiversidad Valenciana (COMAV), Universitat Politècnica de València, Valencia, Valencia, España	MeP6
Claudio Açaí Bracho-Estévanez	Depto. de Biología, IVAGRO, PaisajeLAB – Febimed Group, University of Cádiz, Campus Río San Pedro, Puerto Real. Cádiz, Spain	MeP3

Nombre	Afiliacion	Comunicaciones
Claus Kohfahl	Instituto Geológico y Minero de España (IGME-CSIC), Sevilla, España	MeP7
Cristina Calvo	Universidad de Salamanca	CiP4
Cristina Calvo-López	Universidad de Salamanca	EsO16
Danicka Schröeteller	Departamento de Botánica, Ecología y Fisiología Vegetal. Universidad de Córdoba	MeP5
Daniel Romera Romera	Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba, Córdoba, España	MeO9
David Barrios	Servicio de Bioinformática, Nucleus, Universidad de Salamanca.	EsP3, CiP4
David Barrios	Universidad de Salamanca	EsP3, CiP4
David Barrios-Rogado	Universidad de Salamanca	EsO16
David Durey	Departamento de Frogtek Analytics, Frogtek, Huesca, España	EsO4
David Hervás Marín	Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Valencia, España	SaO4
David Mateos	Agresta S. Coop.	MeP8
David Salgado	Depto. Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España; Depto. Estadística e Investigación Operativa, Universidad Complutense de Madrid, España	CiO5, CiO4
David Salgado	Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España	CiO5, CiO4
Di Pietro, A.	Departamento de Genética, Universidad de Córdoba, Campus Universitario de Rabanales, Edif C5, E-14071 Córdoba, España	MeP1
Diego Nieto Lugilde	Universidad de Cordoba, Córdoba, España	MeO9, DoO11
Diego Nieto Lugilde	Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba	MeO9, DoO11
Eduardo Millan-Ortuondo	Dirección General, Osakidetza SVS.	SaO1
Elena Rosa-Pérez	Subidrección General de Estadísticas Industriales y de Servicios, Instituto Nacional de Estadística, España	CiO4
Elise Dusseldorp	Methodology and Statistics, Leiden University, Leiden, The Netherlands	EsO2
Emilio López Cano	Centro de Investigación para las Tecnologías Inteligentes de la Información y sus Aplicaciones (CETINIA), Universidad Rey Juan Carlos	EsO12

Nombre	Afiliacion	Comunicaciones
Ester Vilaprinyó	Departamento de Ciencias Médicas Básicas, IRBLLIDA, Universitat de Lleida, Lleida, Espanya	EsP2, SaO2
Ester Vilaprinyó	Departamento de Ciencias Medicas Básicas, IRBLLEIDA, Universitat de Lleida, Lleida, España	EsP2, SaO2
Fernando Martinez de Guzmán	Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid	CiO7
Francisco Angel Guerrero Vivas	Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid, España	CiO7
Francisco G. Morillas	Universitat de València, Valencia, España	NAT2
Francisco Goerlich	Departamento de Análisis Económico e Instituto Valenciano de Investigaciones Económicas, Universidad de Valencia, Valencia, España	MeO11
Francisco J Ruiz Gómez	Departamento de Ingeniería Forestal, Universidad de Córdoba	DoO11
Francisco Javier Bonet García	Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba	DoO11
Francisco Jesús Rodríguez Aragón	Oney, Madrid, España	CiO2, EsO10
Francisco Jesús Rodríguez Aragón	Socio R-Hispano, Madrid, España	CiO2, EsO10
Francisco Palmí Perales	Departamento de Economía Aplicada, Universitat de València, València, España	EsO18
Francisco Rodríguez Sánchez	Departamento de Biología Vegetal y Ecología, Universidad de Sevilla, Sevilla, España	MeO4
Fuensanta Arnaldos-García	Departamento de Métodos Cuantitativos para la Economía y la Empresa. Universidad de Murcia	CiP2
Gema Fernández-Avilés	Universidad de Castilla-La Mancha	EsO7
Gemma Pérez-López	Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España	CiO1
Georgina Guilera	Departament de Psicologia Social i Psicologia Quantitativa, Universitat de Barcelona, Barcelona, España.	EsO6
Gonzalo Martínez García	Departamento de Física Aplicada Radiología y Medicina Física, Universidad de Córdoba, Córdoba, España	MeP7
Heidy M.W. den Besten	Food Microbiology, Wageningen University & Research, P.O. Box 17, 6700 AA, Wageningen, the Netherlands	SaO6

Nombre	Afiliacion	Comunicaciones
Inés del Puerto	Universidad de Extremadura, Badajoz, España	EsO15
Inés Garmendia-Navarro	Departamento de Salud, Gobierno Vasco	SaO1
Ismael Lare David	Departamento de Ciencias Agrarias y del Medio Natural, Universidad de Zaragoza, Zaragoza, España	MeP7
Jacqueline Meulman	Mathematical Institute, Leiden University, Leiden, The Netherlands	EsO2
Jaime Ballesteros de la Vega	Unidad de Biometría, Sermes CRO, Madrid, España	CiP1, NAT1
Javier Gomez	URJC, Mostoles, Madrid	MeO3
Javier M. Moguerza	Centro de Investigación para las Tecnologías Inteligentes de la Información y sus Aplicaciones (CETINIA), Universidad Rey Juan Carlos	EsO12
Javier Tarrío-Saavedra	MODES group, CITIC, Departamento de Matemáticas, Escola Politécnica de Enxeñaría de Ferrol, Universidade da Coruña, A Coruña, España	MeO7
Jesús Asín	Departamento de Métodos Estadísticos, Universidad de Zaragoza, Zaragoza, España	MeO2
Jesús López-Fidalgo	Instituto de Ciencia de los Datos e Inteligencia Artificial, Universidad de Navarra, Pamplona, España	EsO1
Jesús Rojo	Departamento de Farmacología, Farmacognosia y Botánica, Universidad Complutense de Madrid, Madrid, España	MeO10
Jesús Sánchez Dávila	Centre de Recerca Ecològica i Aplicacions Forestals (CREAF), Cerdanyola del Vallès, España	MeO4
João Francisco Gonçalves	Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO-InBIO), Universidade do Porto, Vila do Conde, Portugal	MeO12
Jordi Segú Tell	Gerencia de sistemas de información geográfica, Tragsatec, Madrid, Madrid, España	MeO5
Jorge Castillo-Mateo	Departamento de Métodos Estadísticos, Universidad de Zaragoza, Zaragoza, España	MeO2
Jorge Luis Rueda Sánchez	Departamento de Estadística e Investigación Operativa, Universidad de Granada, Granada, España	EsO14
Jorge Sosa	Departamento de Matemática, Escuela Politécnica Nacional, Quito Ecuador	CiO6
Jorge Sosa Donoso	Departamento de Matemáticas, Escuela Politécnica Nacional, Quito, Ecuador	MeO7

Nombre	Afiliacion	Comunicaciones
José-Luis Zafra-Gómez	Departamento de Economía Financiera y Contabilidad, Universidad de Granada, Granada, España	CiO1
José Antonio Martín Fernández	Depto. Informática, Matemática Aplicada y Estadística, Universitat de Girona, Girona	DoP1
Jose Francisco Cobo Díaz	Departamento de Higiene y Tecnología de los Alimentos, Universidad de León, León, España	SaO5
José Jordán Soria	Independent researcher	MeO13
Jose Luis Cañadas Reche	Orange Spain, Madrid, España	EsO5, EsP1
Jose Luis Tomé	Agresta S. Coop.	MeP8
José María Maya-Manzano	Departamento de Biología Vegetal, Ecología y Ciencias de la Tierra, Universidad de Extremadura, Badajoz, España	MeO10
José Mendoza-Bernal y Antonio F. Skarmeta	Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, España	EsO13
José Oteros	Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba, Córdoba, España	MeO10
Jose V. Die	Departamento de Genética, Universidad de Córdoba	DoO11
Juan Alberto Molina Valero	Departamento de Ingeniería Agroforestal, Universidad de Santiago de Compostela, Lugo, España	MeP2
Juan Aparicio Baeza	Instituto Centro de Investigación Operativa (CIO), Universidad Miguel Hernández, Elche, España	EsO3
Juan Claramunt González	Methodology and Statistics, Leiden University, Leiden, The Netherlands	EsO2
Juana Gómez-Benito	Departament de Psicologia Social i Psicologia Quantitativa, Universitat de Barcelona, Barcelona, España.	EsO6
Juana María Alonso Revenga	Departamento de Estadística y Ciencia de los datos, Facultad de Estudios Estadísticos, Universidad Complutense, Madrid, España	DoO10
Julián Rojo	Universidad de Extremadura, Innova-tsn	CiO3
Julio Sandubete	Universidad Francisco de Vitoria, Madrid, España	EsO11
Kah Yen Yeak	Food Microbiology, Wageningen University & Research, P.O. Box 17, 6700 AA, Wageningen, the Netherlands	SaO6
Laura Maldonado-Murciano	Departament de Psicologia Social i Psicologia Quantitativa, Universitat de Barcelona, Barcelona, España.	EsO6
Leonardo Hansa	-	EsO8

Nombre	Afiliacion	Comunicaciones
Licesio J. Rodríguez-Aragón	Escuela de Ingeniería Industrial y Aeroespacial de Toledo, Universidad de Castilla-La Mancha, Toledo, España	EsO1
Lluís Revilla Sancho	Laboratorio de enfermedad inflamatoria intestinal, IDIBAPS, Barcelona, España	DoO12
López-Bergues, M.S.	Departamento de Genética, Universidad de Córdoba, Campus Universitario de Rabanales, Edif C5, E-14071 Córdoba, España	MeP1
López-Díaz, C.	Departamento de Genética, Universidad de Córdoba, Campus Universitario de Rabanales, Edif C5, E-14071 Córdoba, España	MeP1
Lorenzo Escot	Facultad de Estudios Estadísticos, Universidad Complutense de Madrid, España	EsO11
Luis Alberto Rodriguez Ramirez	Universidad Autónoma de Madrid	EsO17
Luis Castro-Martin	Escuela Andaluza de Salud Pública, Granada, España	EsO14
Luis Martinez-Uribe	Fundación Juan March	CiP4
Luis Martínez Uribe	Biblioteca - Centro de apoyo a la investigación. Fundación Juan March, Madrid, España	CiO7
M. Teresa Díaz-Delfa	Departamento de Métodos Cuantitativos para la Economía y la Empresa. Universidad de Murcia	CiP2
M. Victoria Caballero-Pintado	Departamento de Métodos Cuantitativos para la Economía y la Empresa. Universidad de Murcia	CiP2
Mª Ángeles Varo Martínez	Departamento de Ingeniería Forestal, Universidad de Córdoba, Córdoba (España)	MeO1
Maite Barrios	Departament de Psicologia Social i Psicologia Quantitativa, Universitat de Barcelona, Barcelona, España.	EsO6
Marcel H. Zwietering	Food Microbiology, Wageningen University & Research, P.O. Box 17, 6700 AA, Wageningen, the Netherlands	SaO6
María del Mar Rueda	Departamento de Estadística e Investigación Operativa, Universidad de Granada, Granada, España	EsO14
María del Pilar González Barquero	Departamento de Matemáticas, Universidad de Extremadura, Badajoz, España	EsO15
María Durbán	Universidad Carlos-III de Madrid	EsO7
María José Ginzo Villamayor	Santiago de Compostela	CiO8
María Mercedes López Fernández	Departamento de Higiene y Tecnología de los Alimentos, Universidad de León, León, España	SaO5

Nombre	Afiliacion	Comunicaciones
María Neira	Innova-tsn	CiO3
María Victoria Aguilar Pontes	Departamento de Genética, Universidad de Córdoba, Campus Universitario de Rabanales, Edif C5, E-14071 Córdoba, España	MeP1
María Villar Navales	Universidad Complutense de Madrid, Madrid, España	CiP1, NAT1
Mariluz Guillén	Agresta S. Coop.	MeP8
Mercedes Ovejero Bruna	Unidad de Biometría, Sermes CRO, Madrid, España / Departamento de Metodología y Psicobiología, Universidad Complutense de Madrid, Madrid, España	CiP1, NAT1
Mercedes Ovejero Bruna	Unidad de Biometría, Sermes CRO, Madrid, España / Departamento de Metodología y Psicobiología, Universidad Complutense de Madrid, Madrid, España	CiP1, NAT1
Miguel Angel Martinez-Beneito	Departamento de Estadística e Investigación Operativa, Universitat de València, Burjassot, València	EsO18
Miguel Camacho Sánchez	Instituto Andaluz de Investigación y Formación Agraria, Pesquera, Alimentaria y de la Producción Ecológica	MeP4
Miguel Flores	Grupo MODES, SIGTI, Departamento de Matem\'atica, Escuela Politécnica Nacional, Quito Ecuador	CiO6, MeO7
Miguel Flores	MODES group, SIGTI group, Departamento de Matemáticas, Escuela Politécnica Nacional, Quito, Ecuador	CiO6, MeO7
Miguel González	Universidad de Extremadura, Badajoz, España	EsO15
Miriam Esteve Campello	Instituto Centro de Investigación Operativa (CIO), Universidad Miguel Hernández, Elche, España	EsO3
Modesto Escobar	Universidad de Salamanca	CiP4
Modesto Escobar-Mercado	Universidad de Salamanca	EsO16
Olga Vigiak	Joint Reseach Centre, Ispra, Italy	MeO3
Pablo César salazar Zarzosa	Departamento de Botánica, Ecología y Fisiología Vegetal. Universidad de Córdoba	MeP5
Pablo González-Moreno	Departamento de Ingeniería Forestal, Universidad de Córdoba	DoO11, MeP3
Pablo González-Moreno	Department of Forest Engineering, Laboratory of Dendrochronology, Silviculture and Global Change-DendrodatLab- ERSAF, University of Cordoba, Campus de Rabanales. Córdoba, Spain	DoO11, MeP3

Nombre	Afiliacion	Comunicaciones
Palos Fernandez, R.	Departamento de Genética, Universidad de Córdoba, Campus Universitario de Rabanales, Edif C5, E-14071 Córdoba, España	MeP1
Patricia Carracedo Garnateo	Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Valencia, España	SaO4
Patricia Fernández del Valle	Comité de Ética de la Investigación (CEIm) de Cáceres, Fundesalud, Complejo Hospitalario Universitario de Cáceres, Cáceres, España	SaP1
Pedro-José Martínez-Córdoba	Departamento de Administración y Economía de la Empresa, Universidad de Salamanca, Salamanca, España	CiO1
Pedro J. Pérez	Departamento de Análisis Económico, Universitat de València, Valencia, España	NAT2
Pedro Martín-Chávez	Universidad de Extremadura, Badajoz, España	EsO15
Pedro Sandoval	Departamento de Ciencias Médicas Básicas, IRBLLIDA, Universitat de Lleida, Lleida, Espanya	EsP2, SaO2
Pedro Sandoval	Departamento de Ciencias Medicas Básicas, IRBLLEIDA, Universitat de Lleida, Lleida, España	EsP2, SaO2
Pilar Grau-Carles	Facultad de Ciencias Económicas y Empresariales, Universidad Complutense de Madrid, Madrid, España	EsO11
Po-Yu Liu	Department of Internal Medicine, National Taiwan University College of Medicine, Taipei, Taiwan	MeP4
Rafael Villar	Área de Ecología, Dept. de Botánica, Ecología y F. Vegetal, Universidad de Córdoba, Córdoba, España	MeO12
Rafael Villar Montero	Departamento de Botánica, Ecología y Fisiología Vegetal. Universidad de Córdoba	MeP5
Ramón Baiget Llompart	Gerencia de sistemas de información geográfica, Tragsatec, Madrid, Madrid, España	MeO5
Ramón Ferri-Garcia	Departamento de Estadística e Investigación Operativa, Universidad de Granada, Granada, España	EsO14
Roberto Basile	University of L'Aquila	EsO7
Rodriguez López, A.	Departamento de Genética, Universidad de Córdoba, Campus Universitario de Rabanales, Edif C5, E-14071 Córdoba, España	MeP1
Román Mínguez Salido	Universidad de Castilla-La Mancha	EsO7

Nombre	Afiliacion	Comunicaciones
Román Salmerón Gómez	Departamento de Métodos Cuantitativos para la Economía y la Empresa, Universidad de Granada, Granada, España	EsO9
Rosario Martínez-Verdú	Departamento de Economía Aplicad, Universidad de Valencia, Valencia, España	DoO9
Salvador Arenas-Castro	Área de Ecología, Dept. de Botánica, Ecología y F. Vegetal, Universidad de Córdoba, Córdoba, España	MeO12, DoO11
Salvador Arenas-Castro	Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba	MeO12, DoO11
Salvador Naya	MODES group, CITIC, Departamento de Matemáticas, Escola Politécnica de Enxeñaría de Ferrol, Universidade da Coruña, A Coruña, España.	MeO7
Samil Uysal	Methodology and Statistics, Leiden University, Leiden, The Netherlands	EsO2
Sandra Barragán	Departamento de Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España	CiO4, CiO5
Sandra Barragán	Depto. Metodología y Desarrollo de la Producción Estadística, Instituto Nacional de Estadística, España	CiO4, CiO5
Santiago Martin-Malcon	Agresta S. Coop.	MeP8
Saúl Pastor	Servicio de Bioinformática, Nucleus, Universidad de Salamanca.	EsP3
Sergio Jiménez Sanjuán	DNV, Energy Systems	MeO8
Silvia Yepes Barbero	Grado en Administración y Dirección de Empresas (Universidad de Murcia)	CiP3
Tamy Goretty Trujillo-Escobar	Fundación Vasca de innovación e investigación sanitarias Bioef.	SaO1
Valeria Samira Samanamud Taus	Universidad Complutense de Madrid, Madrid, España	CiP1, NAT1
Vicente Coll Serrano	Departamento de Economía Aplicad, Universidad de Valencia, Valencia, España	DoO9
Víctor Javier España Roch	Instituto Centro de Investigación Operativa (CIO), Universidad Miguel Hernández, Elche, España	EsO3
Virgilio Gómez-Rubio	Departamento de Matemáticas, Universidad de Castilla-La Mancha, Albacete España	EsO18
Xavier Barber Vallés	Instituto Centro de Investigación Operativa (CIO), Universidad Miguel Hernández, Elche, España	EsO3