VI JORNADAS CIENTÍFICAS ESTUDIANTES DE LA SEB

14 - 16 SEPT. 2022

UNIVERSITAT DE VALENCIA

FACULTAD DE CIENCAS MATEMÁTICAS
CAMPUS DE BURJASSOT

MÁS INFO.

SEB SOCIEDAD ESPAÑOLA DE BIOESTADÍSTICA

VNIVERSITAT DE VALÈNCIA

Facultad de Ciencias Matemáticas

Departamento de Estadística e Investigación Operativa

VI Jornadas Científicas de Estudiantes de la SEB

Facultad de Ciencias Matemáticas

Campus de Burjassot,
Valencia, España
14, 15 y 16 de septiembre de 2022
http://www.biometricsociety.net/
vi-jornadas-cientificas-de-estudiantes-de-la-seb/

Welcome

¡Bienvenidas y bienvenidos a todas y todos a la VI edición de las Jornadas Científicas de Estudiantes de la Sociedad Española de Bioestadística!

Tras haber tenido que celebrar la última edición en formato online debido a la pandemia mundial provocada por la COVID-19, os comunicamos con mucha ilusión que en la edición de este año podremos encontrarnos de nuevo en Valencia, ciudad en la que se celebraron por primera vez estas jornadas científicas. Como ya sabéis, el objectivo principal de estas jornadas es promover e impulsar el trabajo que los estudiantes y jóvenes investigadores realizamos en el área de la bioestadística. Por esta razón, estas jornadas, de la misma forma que las ediciones anteriores, seguirán siendo un espacio seguro, donde las y los jóvenes estudiantes podremos compartir el trabajo que estamos realizando, sin miedo a críticas o preguntas complejas.

Estas jornadas científicas, organizadas por y para los y las estudiantes de la Sociedad Española de Bioestadística (SEB), se han convertido en una de las actividades principales de la sociedad. No hubiera sido posible organizar ninguna de estas ediciones si no fuera por la ayuda y el apoyo de la SEB, que demuestra año tras año y edición tras edición, que sigue creyendo en el potencial de sus jóvenes estudiantes. Este año no ha sido una excepción, y desde el comité organizador, queremos agradecerle a toda la sociedad el respaldo que nos ha dado, por apoyarnos en todas y cada una de las decisiones que hemos tenido que adoptar y también por la financiación económica.

Queremos agradecer también el apoyo que hemos recibido por parte de la Facultad de Ciencias Matemáticas de la Universitat de València, quien nos ha abierto sus puertas y nos ha ofrecido todas las facilidades, con la intención de que estas jornadas se puedan celebrar en formato presencial. Por otra parte, queremos agradecer también a Joaquín Martínez Minaya por impartir el curso de introducción a inferencia Bayesiana, y a las y los ponentes de la mesa redonda e historia de las jornadas Irantzu Barrio, María José Caballero, David Conesa, Elena Lázaro, Antonio Vicent y Natalia Vilor por aceptar nuestra invitación y participar en estas jornadas.

Para acabar, queremos agradecer a todos los participantes, estudiantes de la SEB, por mostrar interés y confianza en esta sexta edición de las jornadas. ¡Gracias a todas y todos! ¡Esperamos que entre todas y todos, disfrutemos y hagamos que estas jornadas sean especiales!

Comité Científico y Organizador

Scientific and Organizing Committe

- Sofía Aguilar Lacasaña (ISGlobal)
- Gabriel Calvo Bayarri (UV)
- Juan Carbonell Asíns (INCLIVA)
- Martina Cendoya Martínez (IVIA)
- Patricia Genius Serra (BBRC)
- Pavel Hernández Amaro (UC3M)
- Harold Antonio Hernández Roig (UC3M)
- Amaia Iparragirre Letamendia (UPV/EHU)
- Dorota Mlynarczyk (UAB)
- Garazi Retegui Goñi (UPNA)
- Blanca Rodríguez Fernández (BBRC)

Contents

Time schedule

Miércoles 14		Jueves 15		Viernes 16	
		9:00 – 10:00	**Sesión 3:** Estadística espacial		
10:30 – 11:00	Recepción	10:00 – 11:00	**Sesión 4:** Supervivencia	10:00 – 11:15	**Sesión 7:** Genética e Inferencia causal
11:00 – 11:30	Inauguración	11:00 – 11:30	Café	11:15 – 11:45	Café
11:30 – 13:30	Curso	11:30 – 12:30	**Sesión 5:** Machine Learning	11:45 – 13:15	Mesa redonda
		12:30 - 13:30	Pósteres	13:15 – 13:30	Clausura
13:30 – 15:00	Comida	13:30 – 15:00	Comida		
15:00 – 16:00	**Sesión 1**: Software y herramientas	15:00 – 16:30	**Sesión 6:** Biología y ecología		
16:00 – 16:30	Café				
16:30 – 17:45	**Sesión 2**: Salud	16:30 – 17:00	Café		
17:45 – 18:15	Presentación Libro	17:00 – 18:00	Historia de las jornadas		
		19:30 – 21:00	Actividad social/cultural		
		21:00 - ...	Cena		

Miércoles, 14 de septiembre

10:30-11:00 Recepción

11:00-11:30 Inauguración

11:30-13:30 Curso: Be Bayesian my friend. An introduction to Bayesian inference.

13:30-15:00 Comida

SESIÓN 1: SOFTWARE Y HERRAMIENTAS

Chair: Garazi Retegui Goñi

15:00 "Interactive modelling and prediction of patient evolution via multistate models", *Leire Garmendia Bergés.*

15:15 "pluvNav: downloading and dealing with pluviometric data", *Harkaitz Goyena Baroja.*

15:30 "The Coronasurveys project: tracking COVID-19 via indirect reporting", *Harold Antonio Hernández Roig.*

15:45 "`injurytools`: a toolkit for sports injury data analysis", *Lore Zumeta Olaskoaga.*

16:00-16:30 Coffee break

SESIÓN 2: SALUD

Chair: Patricia Genius Serra

16:30 "Desarrollo de un modelo de imputación de datos faltantes. Aplicación en el estudio de la asociación entre la diabetes y la exposición a arsénico.", *María Grau Pérez.*

16:45 "Who is lost to follow-up in the PISCIS Cohort of HIV of Catalonia and Balearic Islands? Analysis of the last 15 years and impact of the COVID-19 pandemic", *Sergio Moreno Fornés.*

17:00 "Flexible dose-response of serum selenoprotein concentrations and proportions by total selenium concentrations in the Aragon Workers Health Study - SelenOmics project.", *Zulema Rodriguez-Hernandez.*

17:15 "Estimación Bayesiana de la validez del dímero D y la escala de Ginebra para el diagnóstico de tromboembolismo pulmonar en pacientes con COVID-19 en la urgencia hospitalaria", *Cristóbal Manuel Rodríguez Leal.*

17:30 "Joint modelling of longitudinally measured body-weight and survival data from the PREDIMED trial", *Andrea Toloba López-Egea.*

17:45-18:15 Presentación Libro

Jueves, 15 de septiembre

SESIÓN 3: ESTADÍSTICA ESPACIAL

Chair: Pavel Hernández Amaro

9:00 ""SurveyMapping": Hacia el análisis geográfico en áreas pequeñas basado en encuestas de salud", *Miguel Ángel Beltrán Sánchez.*

9:15 "Spatial distribution of resistance of *Plurivorosphaerella nawae* to QoI fungicides", *Martina Cendoya Martínez.*

9:30 "Roadmap for assesing statistical modeling of relative biomass indices", *Alba Fuster Alonso.*

9:45 "Multivariate spatio-temporal models for predicting short-term cancer incidence", *Garazi Retegui Goñi.*

SESIÓN 4: SUPERVIVENCIA

Chair: Juan Carbonell Asíns

10:00 "Bayesian Survival Analysis of Acute-On-Chronic Liver Failure in Clinically Stable Outpatients with Cirrhosis", *Pablo Escobar.*

10:15 "A Bayesian spatial illness-death model to assess geographical differences in the risk and incidence of recurrent hip fracture and death", *Fran Llopis-Cardona.*

10:30 "Modelos aditivos y multiplicativos de supervivencia: Un estudio comparado", *Javier Martın-Pozuelo Lozano.*

10:45 "Asociación entre el cáncer de cuello uterino y las medidas antropométricas y la actividad física", *Jon Aritz Panera Carracedo.*

11:00-11:30 Coffee break

SESIÓN 5: MACHINE LEARNING

Chair: Harold Antonio Hernández Roig

11:30 "Machine-learning use of risk prediction models to triage the severity level of COVID-19 patients entering the emergency care system", *Goizalde Badiola-Zabala.*

11:45 "Exploring statistical methods for classifying individuals in extreme aging groups", *Armand González-Escalante.*

12:00 "Development and validation of prognostic models for hospitalization in the Basque Country: Analyzing the variability of non-deterministic algorithms", *Alexander Olza Rodríguez.*

12:15 "Modelos de árboles de regresión y categorización aditivos bayesianos: un encuentro entre dos culturas", *Alfonso Picó.*

SESIÓN DE PÓSTERES

(12:30-13:30)

1. "Modelos de Segmentación Aplicados a la Caracterización del Sector Vacuno Cárnico Español", *María Anciones Polo.*
2. "Aplicación de técnicas estadísticas multivariantes en el análisis, estudio y optimización de los indicadores de lesionabilidad y rendimiento físico en jugadores de fútbol profesional", *Enrique Benéitez Andrés.*
3. "Estudio sobre el microbioma intestinal, y la calidad de vida y el neurodesarrollo en adolescentes", *Raúl Beneyto Menargues.*
4. "Técnicas CHAID Aplicadas a la Caracterización del Sector Vacuno Cárnico Español", *María Anciones Polo.*
5. "Aplicación del Bootstrap", *Gresky Gutiérrez Sánchez.*
6. "Clinical prediction rules for adverse outcomes in patients with SARS COV-2 infection by the omicron variant", *Lander Rodríguez Idiazabal.*
7. "The food traffic light that gives a green light to ultra-processed foods: visual data mining", *María Carmen Romero Ferreiro.*

13:30-15:00 Comida

SESIÓN 6: BIOLOGÍA Y ECOLOGÍA

Chair: Martina Cendoya Martínez

15:15 "Analysis of longitudinal Microbiota data using a Dirichlet Autoregressive Model", *Irene Creus-Martí.*

15:30 "Records tests and applications to climate change", *Jorge Castillo-Mateo.*

15:45 "Evaluación de la idoneidad climática de la cuenca Mediterránea para el desarrollo de la mancha negra de los cítricos, causada por *Phyllosticta citricarpa*", *Anaïs Galvañ.*

16:00 "Forecast of temperature-attributable mortality at lead times of up to 15 days for a very large ensemble of European regions", *Marcos Quijal-Zamorano.*

16:15 "Biplot logistico asociado al análisis de la redundancia para datos de respuesta binaria", *Laura Vicente González.*

16:30-17:00 Coffee break

17:00-18:00 Historia de las jornadas

19:30-21:00 Actividad social/cultural

21:00 Cena

Viernes, 16 de septiembre

SESIÓN 7: GENÉTICA E INFERENCIA CAUSAL

Chair: Sofía Aguilar y Blanca Rodríguez

10:00 "Epigenome-wide study of the exposure to green spaces and blood DNA methylation", *Sofía Aguilar Lacasaña.*

10:15 "Renyi divergence measures for the evaluation of surrogate endpoints based on causal inference", *Gokce Deliorman.*

10:30 "Exploring quantitative brain features associated with high genetic predisposition to Alzheimer's disease using Compositional Data Analysis ", *Patricia Genius Serra.*

10:45 "Penalized Logistic Regression for Health Status Classification Using Gene Expressions ", *Carlos Javier Peña de los Santos.*

11:00 "A comparison of Mendelian Randomization methods for assessing causal effects on complex traits", *Blanca Rodríguez Fernández.*

11:15-11:45 Coffee break

11:45-13:15 Mesa redonda: "El papel transversal de la estadística en la investigación científica "

13:15-13:30 Clausura

Curso

Title: Be Bayesian my friend. An introduction to Bayesian inference.

Instructor: Joaquín Martínez Minaya, Universitat Politècnica de València (UPV)

Abstract: Bayesian inference is becoming increasingly popular in many data analysis fields, also in disciplines such as epidemiology, ecology, industry, economy or medicine. This way of doing inference was born as a consequence of Bayes' Theorem. In this course, we will see how to combine the main elements of the Bayesian inference, the prior distribution and likelihood, with the aim of obtaining through the Bayes' Theorem the posterior and predictive distribution. We will also introduce the hierarchical Bayesian models, how them should be constructed, and how can we fit them using Markov Chain Monte Carlo (MCMC) techniques and the Integrated Nested Laplace Approximation (INLA). Lastly, we will apply this knowledge to some real examples.

Historia de las jornadas

Título: ¿De dónde venimos y hacia dónde vamos? Historia de las jornadas.

Resumen: Las Jornadas Científicas de Estudiantes de la SEB llegan a su sexta edición, y vuelven al origen, a donde comenzó todo, a Valencia. En esta sesión especial dedicada a la historia de las jornadas haremos un repaso de la evolución de las mismas con datos, fotos y gráficos, desde sus comienzos con la primera edición celebrada en enero del 2015, hasta la sexta edición que se va a celebrar en septiembre del 2022. Para ello, contaremos con tres ponentes de gran relevancia para las jornadas por ser los impulsores y creadores de la primera edición: **David Conesa**, Profesor Catedrático de la Universitat de Valencia (UV) y presidente de la SEB en la primera edición de las jornadas; **Irantzu Barrio**, Profesora de la Universidad del País Vasco (UPV/EHU), actual tesorera de la SEB y organizadora de la primera edición de las jornadas; y **Natalia Vilor-Tejedor**, investigadora postdoctoral en el Center for Genomic Regulation (CRG) - Barcelonabeta Brain Research Center (BBRC) y vocal de la actual junta directiva de la SEB.

Mesa Redonda

Título: "El papel transversal de la estadística en la investigación científica"

Resumen: Si hay una ciencia transversal que esté en constante conversación y contacto con el resto de ciencias empíricas, ésta es, sin duda, la estadística. Está presente tanto en las ciencias naturales como en las ciencias sociales y es utilizada como herramienta para producir conocimiento. Aquellas personas dedicadas a la investigación científica pueden encontrarse con dificultades para entenderla o, si son conocedoras de la materia, para transmitirla. Para hacernos una idea clara de esta situación contaremos con la participación de 4 panelistas invitados: María José Caballero (GVA), Elena Lázaro Hervás (IVIA), David Conesa Guillén (UV) y Antonio Vicent Civera (IVIA). Gracias a sus experiencias y conocimientos podremos conocer cuestiones vitales para un investigador tales como: cuán importante es la estadística en otras ramas de la investigación; cuánta importancia se le otorga a la rigurosidad a la hora de transmitir conocimientos y/o publicar artículos; cuáles son las mayores diferencias para un investigador entre el mundo empresarial y el mundo académico. Estas son solo algunas cuestiones que abordaremos, además se incitará al debate entre los panelistas sobre cualquier experiencia interesante que puedan compartir. También contaremos con rondas de preguntas por parte del público, así que no te quedes con ninguna duda.

María José Caballero: Generalitat Valenciana (GVA) María José Caballero es matemática y egresada del Máster en Bioestadística por la Universidad de Valencia. Actualmente trabaja en el Servicio de Salud digital y espacio de datos en la Dirección general de Planificación, Eficiencia Tecnológica y Atención al Paciente de la Conselleria de Sanidad, dedicación compartida con la docencia como profesora asociada en el Departamento de Estadística e Investigación Operativa de la Universidad de Valencia. Su carrera profesional ha estado siempre enfocada en el área de la salud, desarrollándose como investigadora en diferentes proyectos nacionales y europeos a la par que profundizando sus conocimientos en los diferentes sistemas de información de salud de la Comunidad Valenciana. Además de su labor investigadora en varios Institutos de Investigación, tuvo la oportunidad de ser la responsable del Servicio de Estadística en FISABIO. Todas estas líneas de trabajo le han permitido ejercer la profesión de estadística en equipos multidisciplinares con profesionales de diferentes perfiles, donde esta ciencia, en ocasiones, no es una gran conocida.

David Conesa Guillén: Universitat de València (UV). David Conesa es Catedrático de Estadística e Investigación Operativa en la Universitat de València, donde trabaja desde octubre de 1993 y en la que ejerce en la actualidad de director del departamento de Estadística e Investigación Operativa. Desarrolla investigación en el área de la modelización estadística de situaciones en las que la incertidumbre está presente, y lo hace mayoritariamente desde la perspectiva bayesiana. Así, por ejemplo, ha trabajado en problemas de modelos de colas de espera de trasplantes, modelos de detección de umbrales de enfermedades, análisis de eficiencia, modelos de supervivencia animal, y últimamente en modelos de distribución espacial de especies y enfermedades. Fue el Presidente de la Sociedad Española de Bioestadística durante los años 2014 y 2015, y en la actualidad es el Editor en Jefe de la revista Statistics and Operations Research Transactions (SORT).

Elena Lázaro Hervás: Valencian Institute for Agricultural Research (IVIA). Elena Lázaro es Ingeniera Agrónoma por la Universitat Politècnica de València, Máster en Bioestadística y Doctora en Estadística y Optimización por la Universitat de València. Actualmente trabaja como bioestadística en la Unidad de Micología del Departamento de Protección Vegetal y Biotecnología del Instituto Valenciano de Investigaciones Agrarias (IVIA) y colabora como experta con la Autoridad Europea de Seguridad Alimentaria (EFSA) en cuestiones de vigilancia epidemiológica. Esta especialización profesional ha hecho que su carrera investigadora se centre en la aplicación y el desarrollo de métodos estadísticos en diferentes contextos "bio" más allá de la investigación agraria. Actualmente su trabajo se centra en el ámbito de la sanidad vegetal, principalmente en el desarrollo y aplicación de modelos epidemiológicos para patógenos cuarentenarios con vistas a la mejora de la eficacia de sus estrategias de vigilancia y de gestión de riesgos.

Antonio Vicent: Valencian Institute for Agricultural Research (IVIA). Doctor Ingeniero Agrónomo por la Universidad Politécnica de Valencia y Máster en Bioestadística por la Universidad de Valencia. Inicia su carrera investigadora en 1998 como becario en la Unidad de Patología Vegetal de la ETSI Agrónomos de Valencia, donde desarrolla su actividad hasta el año 2009, cuando se incorpora como investigador en la Unidad de Micología del IVIA, de la que actualmente es el responsable. Desde 2021 es también Coordinador del Centro de Protección Vegetal y Biotecnología del IVIA. Su línea de trabajo se centra en el desarrollo de programas de control integrado de enfermedades causadas por hongos, principalmente en cítricos y otros frutales subtropicales. En esta área ha liderado varios proyectos de investigación del plan nacional y convenios con empresas. Realiza también estudios de análisis de riesgos de introducción de enfermedades exóticas y de cuarentena en la UE. Ha participado en varios grupos de trabajo internacionales y proyectos europeos sobre patógenos de cuarentena. Actualmente es vocal de la Sociedad Española de Fitopatología y miembro de Panel de Sanidad Vegetal de la Autoridad Europea de Seguridad Alimentaria (EFSA).

Speakers

Sesión 1: Software y herramientas
14 de septiembre, 15:00 a 16:00

Chair: Garazi Retegui Goñi

Sep 14th
15:00

Interactive modelling and prediction of patient evolution via multistate models

Leire Garmendia Bergés[1], Jordi Cortés Martínez[1], Guadalupe Gómez Melis[1]

[1]Department of Statistics and Operational Research, Universitat Politécnica de Catalunya

Modelling the disease course regarding serious events and identifying prognostic factors is of great clinical relevance. Previous studies to predict high-risk critically ill cases among COVID-19 hospitalized patients have not yet arrived at a solid conclusion. Besides death, other intermediate events such as the need for invasive ventilation are relevant for clinical management.

A team formed by clinicians and biostatisticians worked on the identification of the most clinically relevant states to explain the evolution of COVID-19 hospitalized patients, on the meaningful and plausible transitions between them, and on the characterization of the prognostic factors for those states. Based on this consensus, a multistate model (MSM) is proposed in order to learn about the disease progress.

Motivated by this situation, an app is presented with two main goals: 1) to fit a MSM from specific data in a friendly way; 2) to predict the clinical evolution for a given patient based on the previous MSM. For the first objective, the user defines the states and transitions of the model as well as the covariates involved in each transition. The app returns descriptive information through histograms or barplots for the covariates, by box-plots to show the length of stay for each state and through instantaneous hazard plots to represent the risk of transition over time. Regarding the model, the app returns the estimated coefficients as well as the predictive performance measured by the logarithmic score. For the second goal, information of the new patient at an initial state such as age or sex and the time for which predictions want to be made has to be provided. From these inputs, the app provides some indicators of the patient's evolution such as the probability of each state at a fixed time. Furthermore, visual representations (e.g., the stacked transition probabilities plot) are given to make predictions more understandable.

For illustrative purposes, we show how the app works using data from a multicohort study of more than 5,000 hospitalized adult COVID-19 patients from 8 Catalan hospitals during the first five waves of the pandemic. Different models have been fitted for the first Catalan pandemic wave, including as states the main outcomes –discharge and death– together with objective interventions during hospitalization such as non-invasive or invasive mechanical ventilation.

The application and the underlying model are intended to be very useful for clinicians and to enhance the approach in modelling the course of other diseases with different stages of severity.

Keywords: Multistate model, shiny app, COVID-19

---

pluvNav: downloading and dealing with pluviometric data

Sep 14th
15:15

Harkaitz Goyena[1,2], Unai Pérez-Goya[1,2], Ana F. Militino[1,2]

[1] Department of Statistics,Computer Science and Mathematics, Public University of Navarre, Pamplona, Spain; [2] INAMAT[2] Institute, Pamplona, Spain;

Precipitation, temperature, wind, and pollution are directly related to atmospheric and climatic changes. These changes have a cascade effect on other fields such as public health or economy. Access to information captured by public and private organisations is nowadays a key asset in the study of different fields. European governments are currently making an effort to publish these observations in a public and open way. However, there is no standardized platform to obtain data from all the different sources, e.g. in Spain some regions have additional rainfall stations that only publish their data in a regional database. Navarre is an example of this issue.

The aim of the `pluvNav` package is to provide automatic tools to download and process rainfall data for the Spanish province of Navarre for the `R` programming language. The package downloads all the observations of two types of rainfall stations: automatic and manual. The package also allows for the unification of data into text files ready to use for modelling and forecasting using `R` language. In addition, the package allows the user to fill in some missing values or measurements to obtain more suitable datasets. Users can run the whole downloading and filling process in a convenient way using a series of `R` commands.

Keywords: Pluviometry, Data Collection

Sep 14th
15:30

The Coronasurveys project: tracking COVID-19 via indirect reporting

Harold A. Hernández-Roig[1,2], Antonio Fernández Anta[3], Rosa E. Lillo[1,2], Coronasurveys Team[4]

[1] Department of Statistics, Universidad Carlos III de Madrid; [2] UC3M-Santander Big Data Institute; [3] IMDEA Networks Institute; [4] https://coronasurveys.org/team/.

Coronasurveys (https://coronasurveys.org/) is a collaborative platform between several universities, research institutions, and independent volunteers. It collects data about the COVID-19 pandemic using indirect reporting via open surveys. Particularly, it relies on the Network Scale-up Method (NSUM) to provide multiple estimates related to the evolution of the pandemic, like the number of infected, fatalities, active cases, etc. These estimates are available for a large number of countries and their evolution over time is been recorded since March 2020. In combination with other data sources, the platform provides a simple, cheap, and homogeneous tracking tool that is freely available for researchers, decision-makers, and citizens in general.

In this talk, we describe the Coronasurveys system for data collection and processing, as discussed in [1]. Furthermore, we describe the results of our efforts when estimating the COVID-19 prevalence in Spain using the NSUM [2]. We will also discuss several forecasting methods that have been deployed for our surveys and other data sources that have been incorporated into our database. Finally, we will discuss some of the future research lines and challenges that are still to be tackled.

Keywords: COVID-19, Open surveys, Network Scale-up Method.

References

[1] Baquero, C., Casari, P., Fernandez Anta, A., García-García, A., Frey, D., Garcia-Agundez, A., Georgiou, C., Girault, B., Ortega, A., Goessens, M., Hernández-Roig, H. A., Nicolaou, N., Stavrakis, E., Ojo, O., Roberts, J. C., and Sanchez, I. (2021). The CoronaSurveys System for COVID-19 Incidence Data Collection and Processing. *Frontiers in Computer Science*, 3(June), 1–10.

[2] Garcia-Agundez, A., Ojo, O., Hernández-Roig, H. A., Baquero, C., Frey, D., Georgiou, C., Goessens, M., Lillo, R. E., Menezes, R., Nicolaou, N., Ortega, A., Stavrakis, E., and Fernandez Anta, A. (2021). Estimating the COVID-19 Prevalence in Spain With Indirect Reporting via Open Surveys. *Frontiers in Public Health*, 9(April), 1–5.

`injurytools`: a toolkit for sports injury data analysis

Sep 14th
15:45

Lore Zumeta-Olaskoaga[1,2], Dae-Jin Lee[2]

[1]BCAM - Basque Center for Applied Mathematics;
[2] Departamento de Matemáticas, Universidad del País Vasco UPV/EHU

We present an **R** package that provides handy tools to analyze sports injury data. Currently, there exist consensus procedures for injury data collection [1], plus more and more information, that could be related to injuries, are being collected thanks to the advance of new technologies. But the collection of data is undoubtedly useless without exploiting and extracting relevant information from them. And it should further be noted that sports injury data encompasses some particularities. In this context, this package provides functions to facilitate: sports injury data preparation, visualization and estimation.

Thereby, the first and main step involves transforming the data in an adequate structure so that it is prepared for use in statistical analyses. A key data entry here is the time of exposure, i.e., the period of time over which players or athletes have been exposed to the risk of injury, measured as one of: minutes training and/or participating in competition events (e.g. matches), hours, days or seasons. Then, taking into account the time of exposure, common statistical summaries can be calculated, such as epidemiological measures of injuries: incidences and injury burdens [1, 2]. Moreover, several convenience visualization functions are available. In practice, this can help automate certain descriptive reports that are routinely made for sports injury surveillance. The estimation of the risk of injury with other covariate effects is performed outside of `injurytools`, whether the event of injury (outcome variable) is seen as count or time-to-event data.

In this talk, we show an exemplary sports injury data workflow using convenience functions from `injurytools`.

Keywords: software, sports injury prevention research, epidemiology

References

[**1**] Fuller, C. W., Ekstrand, J., Junge, A., Andersen, T. E., Bahr, R., Dvorak, J., ... & Meeuwisse, W. H. (2006). Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *Scandinavian journal of medicine & science in sports*, 16(2), 83-92.

[**2**] Bahr, R., Clarsen, B., & Ekstrand, J. (2018). Why we should focus on the burden of injuries and illnesses, not just their incidence. *British Journal of Sports Medicine*, 52(16), 1018-1021.

Sesión 2: Salud
14 de septiembre, 16:30 a 17:45

Chair: Patricia Genius Serra

Desarrollo de un modelo de imputación de datos faltantes. Aplicación en el estudio de la asociación entre la diabetes y la exposición a arsénico.

Sep 14th 16:30

María Grau Pérez[1,2,3], Josep Redon[1], Jose Luis Gómez Ariza[4], Tamara García Barrera[4], Juan C. Martín-Escudero[5], José Bermúdez Edo[3], María Téllez Plaz[2,6]

[1]Instituto de Investigación Biomédica INCLIVA, Valencia; [2]Universidad Autónoma de Madrid, Madrid;[3]Universidad de Valencia, Valencia; [4]Universidad de Huelva, Huelva; [5]Hospital Universitario Rio Hortega, Universidad de Valladolid, Valladolid; [6]Centro Nacional de Epidemiología, Institutos de Salud Carlos III, Madrid

Introducción: La imputación es una técnica tradicional para el tratamiento de datos faltantes. La imputación múltiple (IM) consiste en asignar a cada dato no observado m>1 valores imputados. En este trabajo se propone un método de IM que será aplicado para estudiar la asociación entre la exposición al arsénico y la prevalencia de diabetes.

Métodos: Seleccionamos 1451 participantes del estudio Hortega, una muestra representativa de la población general de Valladolid. Todos los participantes tienen mediciones de arsénico total en orina, mientras que las mediciones de arsenobetaína, una especie de arsénico orgánico no tóxica para el ser humano, se llevó a cabo en una submuestra aleatoria de 295 participantes, dejando así un 79.7% de valores faltantes. Para corregir por la confusión que introduce el arsénico orgánico en las mediciones de arsénico total, desarrollamos un método de IM utilizando cadenas de Markov. En particular, generamos m=5 simulaciones de los valores no medidos de arsenobetaína, generando 5 bases de datos sin valores faltantes. Con cada base de datos completa se usan modelos logísticos multivariantes para estudiar la asociación entre el arsénico total medido en orina y la prevalencia de diabetes. Finalmente, los m=5 resultados obtenidos se combinan en una única solución mediante la teoría de IM propuesta por Rubin. El modelo desarrollado es validado utilizando una muestra independiente.

Resultados: 108 participantes presentan prevalencia de diabetes. El odds ratio (intervalo de confianza 95%) de diabetes comparando un incremento en los niveles de arsénico total del percentil 20 al percentil 80 fue 1.31 (0.94, 1.81) antes de ajustar por arsenobetaína, y 1.76 (1.54, 2.02) después de ajustar por arsenobetaína. El método de IM converge correctamente y no muestra sensibilidad hacia un cambio en las distribuciones previas proporcionadas. La validación del modelo muestra que las imputaciones realizadas son apropiadas incluso con el 80% de datos faltantes imputados.

Conclusión: La exposición a arsénico se asocia con una mayor prevalencia de diabetes. Es importante ajustar por arsenobetaína cuando queremos estimar los efectos de la exposición a arsénico en poblaciones con un consumo alto de pescado. El método de IM por cadenas de Markov es apropiado para ser aplicado en otros estudios con problemas de datos faltantes, incluso con tasas de no respuesta elevadas.

Keywords: missing data, multiple imputation, epidemiology

Who is lost to follow-up in the PISCIS Cohort of HIV of Catalonia and Balearic Islands? Analysis of the last 15 years and impact of the COVID-19 pandemic.

Sep 14th
16:45

Sergio Moreno-Fornés[1,2,3,*], Jorge Palacio-Vieira, Yesika Díaz, Jordi Aceitón, Andreu Bruguera, Daniel K. Nomah, Josep M. Llibre, Hernando Knobel, Iván Chivite, José María Miro, Jordi Casabona, Arkaitz Imaz, Juliana Reyes-Urueña and PISCIS study group.

[1]Centre Estudis Epidemiològics sobre les Infeccions de Transmissió Sexual i Sida de Catalunya (CEEISCAT), Dept Salut, Generalitat de Catalunya, Badalona, Spain; [2]CIBER Epidemiologia y Salud Pública (CIBERESP), Barcelona, Spain; [3]Institut d'Investigació Germans Trias i Pujol (IGTP), Barcelona, Spain

Background: Retaining people living with HIV (PLWH) in care and ensuring good adherence to antiretroviral therapy (ART) remain as cornerstones to achieving the UNAIDS 95-95-95 targets. Few studies have described how the profile of PLWH lost to follow-up (LTFU) evolved over time. We aimed to describe this evolution by grouping demographic and clinical characteristics of individuals LTFU from 2006 to 2020 and to analyze the impact of COVID-19 on the number of people LTFU. **Methods:** The PISCIS Cohort follows PLWH in sixteen hospitals of Catalonia and two of the Balearic Islands. Patients were considered LTFU in each particular year if they had no contact with the HIV clinic for $\geq$12 months. Latent class (LC) analysis was used to identify subgroups of PLWH lost to follow-up with similar clinical and epidemiological characteristics by year. The number of LC for each year were chosen according to statistical results (BIC, aBIC, cAIC, Entropy > 0.8 and all classes having more than 5% of population) as well as clinical criteria. A longitudinal mixed model was calculated to analyze the impact of COVID-19 on the magnitude of LTFU by comparing data from 2018-2019 with data from 2020. **Results:** A total of 2,841 patients (14.58%) out of 19,481 were LFTU. The prevalence of LTFU decreased from 6.2% in 2006 to 0.8% in 2019, with a slight increase in 2020 (1.1%), changing the causes over time. Eight latent classes of individuals LTFU were identified (Figure 1): two groups of people who inject drugs (PWID), three LC of men who have sex with men (MSM), two groups of women and only one LC of heterosexual men. Most of the LC were selected by the entropy criteria, followed by the lowest value of BIC, cAIC and aBIC respectively. Conversely, individuals aged between 30 to 64 years and having lived 10 or more years with HIV were less likely to be LTFU during the pandemic. Compared to 2018-2019, during the COVID-19 pandemic, the proportion of LTFU was higher among women (15.5% vs. 20.6%), but without statistical significance in the multivariate logistic model. **Conclusions:** Characteristics of LTFU changed over time, as most men LTFU were initially PWID and at the end were MSM. The COVID-19 pandemic has minor impact on LFTU, and it is observed among older patients and those living longer with HIV. The analysis of LTFU overtime could be useful to tailor HIV prevention strategies.

Keywords: HIV, loss of follow-up, retain to care.

Flexible dose-response of serum selenoprotein concentrations and proportions by total selenium concentrations in the Aragon Workers Health Study - SelenOmics project.

Sep 14th
17:00

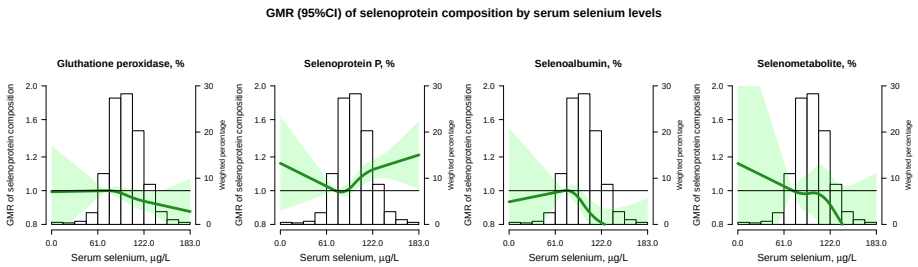
Zulema Rodriguez-Hernandez[1,2,*], Anabel Paredes-Douton[1,*], Marta Galvez-Fernandez[3,*], Maria Grau-Perez[4], Jose L. Gomez-Ariza[5], Tamara Garcia-Barrera[5], Belén Callejón-Leblic[5], Martin Laclaustra-Gimeno[6], Belen Moreno-Franco[6], Fernando Civeira[6], José Puzo[6], Jose A. Casasnovas[6], Roberto Pastor-Barriuso[1] and Maria Tellez-Plaza[1]

[1]National Center for Epidemiology, Carlos III Health Institutes; [2]Department of Biotechnology, Universitat Politècnica de València; [3]Department of Environmental Health Sciences, Columbia University; [4]Biomedical Research Institute INCLIVA; [5]Department of Chemistry, University of Huelva; [6]IIS Aragon, CIBERCV, Zaragoza University. *Equal author contribution.

Background. Selenium (Se) deficiency has been related to adverse health effects, because Se is a key component of antioxidant selenoproteins. Excess of selenium, however, also has a negative impact on human health, as it could lead to a non-specific binding of Se beyond selenoproteins and increased levels of bioactive Se metabolites. However, population-based samples with available selenium speciation data are rare.

Materials and methods. Serum selenoproteins (glutathione peroxidase [GPx], selenoprotein P [SeP], selenoalbumin [SeAlb]), Selenometabolites and total Se were measured using HPLC/ICP-MS in 862 AWHS participants. We used restricted quadratic splines to evaluate the flexible dose-response of serum selenoproteins and selenometabolites (as concentration and as proportion, out of the sum of species) by serum total Se, a biomarker of Se status, in the Aragon Workers Health Study.

Results. Serum Se levels were linearly associated with increased concentrations of all selenoproteins, but not selenometabolites (data not shown). At total Se higher than ∼110 μg/L, we observed an inverse association with the logit-transformed proportion of GPx, SeAlb and Se-metabolites, which are known to be more efficiently eliminated under chronic excessive exposure conditions, and a positive association with SeP, which is the blood transporter protein to other tissues.



Conclusions. Results are compatible with a saturation of antioxidant selenoproteins at high Se exposure. Future compositional data analysis is needed to understand how selenoproteins and selenometabolites composition influence specific health endpoints.

Keywords: selenium, selenoproteins, flexible dose-response

Estimación Bayesiana de la validez del dímero D y la escala de Ginebra para el diagnóstico de tromboembolismo pulmonar en pacientes con COVID-19 en la urgencia hospitalaria

Sep 14th
17:15

Rodríguez-Leal CM[1], Susi-García MR[2], Amador-Pacheco J[2]

[1]Servicio de Urgencias, Hospital del Henares, Coslada (Madrid); [2]Departamento de Estadística y Ciencia de los Datos, Facultad de Estadística, Universidad Complutense de Madrid

Introducción. El tromboembolismo pulmonar (TEP) es una posible complicación de los pacientes con COVID-19, cuya frecuencia de aparición está aumentada en ellos respecto a población general. Su diagnóstico no es sencillo y existen recursos como la escala de Ginebra y el dímero D (DD) que pueden ayudar al mismo, pero no están validados en pacientes infectados por SARS-CoV-2. Para el diagnóstico de TEP se utiliza la angiografía pulmonar mediante tomografía computarizada (TC-TEP), que constituye un *gold standard* imperfecto. La estadística bayesiana proporciona el marco teórico necesario para poder evaluar la validez de pruebas diagnósticas en ausencia de una prueba perfecta de confirmación de referencia. **Objetivo.** Evaluar la validez de la escala de Ginebra y el DD en el diagnóstico de TEP en paciente con COVID-19 en la Urgencia Hospitalaria. **Material y métodos.** Se realizó una estimación del área bajo la curva (AUC) de la característica operativa del receptor (curva ROC) mediante metodología bayesiana, y se comparó el resultado con la aproximación empírica. Se seleccionó una muestra de pacientes consecutivos con COVID-19 atendidos en la Urgencia del Hospital Universitario del Henares (Madrid) durante los meses de marzo a diciembre de los años 2020 y 2021, en los que se sospechó TEP, se realizó una TC-TEP y se determinó el DD. La muestra piloto la constituyeron pacientes de las mismas características atendidos en enero y febrero de 2021, y a partir de la misma se elicitó la información a priori. **Resultados.** Se analizaron 45 pacientes con diagnóstico radiológico de TEP y 131 sin él. Fue precisa una transformación de Box-Cox para que la distribución muestral de los datos se ajustara a una distribución de probabilidad gamma y se aplicó un modelo bigamma. Los resultados más precisos los proporcionó el modelo con *gold standard* y distribución de probabilidad a priori informativa, en tanto que los modelos con distribución de probabilidad a priori no informativa y/o sin *gold standard* presentaron problemas de convergencia y un peor ajuste a los datos. Los resultados obtenidos fueron consistentes entre las diferentes técnicas estadísticas utilizadas. **Conclusión.** La escala de Ginebra no es útil y el dímero D sí lo es. Con esta prueba diagnóstica, un umbral en torno a 700 ng/mL parece apropiado como punto de corte de cribado para descartar TEP en estos pacientes. Los modelos bayesianos son los únicos posibles en ausencia de *gold standard*, pero si la prueba no es muy buena surgen problemas de identificabilidad con peor exactitud y convergencia. El uso de una distribución de probabilidad *a priori* informativa mejora la convergencia, pudiéndose incluso superar violaciones de la distribución de probabilidad para un modelo paramétrico. Si es posible, es mejor usar un modelo en presencia de *gold standard*, sobre todo para pruebas modestas.

Keywords: Curva ROC, Embolia Pulmonar, Infecciones por Coronavirus

Joint modelling of longitudinally measured body-weight and survival data from the PREDIMED trial

Sep 14th
17:30

Andrea Toloba[1], Klaus Langohr[1], Guadalupe Gómez[1], Isaac Subirana[2]

[1]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya; [2]CIBER of Epidemiology and Public Health (CIBERESP), Instituto de Salud Carlos III

Overweight and obesity are associated with higher risk of cardiovascular mortality and all-cause mortality. Weight loss is commonly prescribed as a lifestyle intervention for these patients. However, weight loss is frequently followed by repeated episodes of subsequent regain of the lost weight, resulting in weight fluctuation. Whether such fluctuations in body weight are associated with worse prognosis is controversial [1], leading to questions about the prudence of recommending weight loss in obese patients.

We aim to explore that association with data from the PREDIMED trial, which provides annual body-weight measurements of 7,447 subjects followed up until death. Previous research studies on this matter were based on empirical definitions of body-weight fluctuation, such as the difference between successive recorded values or the variance of all intra-individual recorded weights. Instead, we propose to fit a joint model of longitudinal and time-to-event data. Weight over time is adjusted with a linear mixed model, and time-to-death is modelled through a proportional hazards model. Shared random effects account for the dependency between longitudinal dropout and survival status, and an appropriate link function is set to capture body-weight fluctuation. Parameter estimation is achieved with function `jointModel()` from the `JM` [2] package.

Keywords: Joint models, Linear mixed models, Survival analysis

References

[1] Zou H., Yin P., Liu L., Liu W., Zhang Z., Yang Y., Li W., Zong Q., and Yu X. (2019). Body-Weight Fluctuation Was Associated With Increased Risk for Cardiovascular Disease, All-Cause and Cardiovascular Mortality: A Systematic Review and Meta-Analysis. *Frontiers in Endocrinology*, 10, 728.

[2] Rizopoulos D. (2010). JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. *Journal of Statistical Software*, 35(9), 1–33.

Sesión 3: Estadística espacial
15 de septiembre, 9:00 a 10:00
Chair: Pavel Hernández Amaro

"SurveyMapping": Hacia el análisis geográfico en áreas pequeñas basado en encuestas de salud

Sep 15th 09:00

Miguel Ángel Beltrán Sánchez[1], Miguel Ángel Martínez Beneito[2], Paloma Botella Rocamora[3], Jordi Pérez Panadés[3], Francisca Corpas Burgos[3]

[1]Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (Fisabio); [2] Departament d'Estadística i Investigació Operativa. Facultat de Ciències Matemàtiques, Universitat de València; [3]Conselleria de Sanitat Universal i Salut Pública. Generalitat Valenciana

Los estudios geográficos en áreas pequeñas suponen una excelente herramienta epidemiológica. La mayoría de estudios de este tipo se han utilizado para monitorizar problemas de salud a partir de eventos concretos, en particular de conteos de muertes [1]. Habitualmente, estos estudios se basan en el análisis de la información de registros o bases de datos sanitarias en general. Sin embargo, el ceñir los estudios geográficos únicamente a estas fuentes supone una gran limitación. La explotación de otras fuentes alternativas, como por ejemplo las Encuestas de Salud, supone una ampliación del campo de aplicación de los estudios en áreas pequeñas y permite explorar otros indicadores de salud que el estudio de bases de datos sanitarias no permite abordar. Sin embargo, el complejo diseño estadístico de muchas encuestas, en concreto de las Encuestas de Salud, hace que los métodos de análisis en áreas pequeñas no sean de aplicación directa al análisis de encuestas. En este trabajo se adaptan los métodos de análisis en áreas pequeñas al análisis de datos de encuestas. La metodología a utilizar se basa en los modelos jerárquicos bayesianos. Aplicamos dichos modelos al análisis de la Encuesta de Salud de la Comunidad Valenciana (CV) para describir la distribución geográfica de distintos indicadores de interés para la salud en esta región.

Keywords: Mapeo de enfermedades, análisis de encuestas, análisis de datos ordinales

References

[1] Martínez Beneito, M. A. and Botella Rocamora, P. (2019). Disease Mapping: From Foundations to Multidimensional Modeling. *CRC Press*.

Sep 15th
09:15

Spatial distribution of resistance of *Plurivorosphaerella nawae* to QoI fungicides

Martina Cendoya[1], Luan Vitor Nascimento[1,2], David Conesa[3], Josep Armengol[2], Mónica Berbegal[2], Antonio Vicent[1]

[1]Centre de Protecció Vegetal i Biotecnologia, Institut Valencià d'Investigacions Agràries (IVIA); [2]Instituto Agroforestal Mediterráneo, Universitat Politècnica de València; [3]Valencia Bayesian Research Group (VaBaR), Departament d'Estadística i Investigació Operativa, Universitat de València

The persimmon tree disease known as circular leaf spot, caused by the fungus *Plurivorosphaerella nawae*, induces leaf necrosis, defoliation and fruit drop. The disease has been effectively controlled with fungicide sprays that include QoI (strobirulins) among others. A lack of efficacy of these applications was observed in 2019 in some persimmon growing areas in the Comunitat Valenciana. In 2020, the main persimmon production area was surveyed, and 183 isolates of *P. nawae* were obtained from 60 georeferenced locations (plots). The resistance of the isolates to QoIs fungicides was analyzed phenotypically and molecularly. The resistance was related to the high prevalence of the G143A mutation in the cytochrome b gene, found in 55.4% of the isolates analyzed. Isolates with this mutation had also lower sensitivity to pyraclostrobin (QoI fungicide) than those without the mutation.

The distribution of the occurrence of resistant isolates in the study area was analyzed by means of spatial Bayesian hierarchical model through the INLA methodolgy [2]. Due to the heterogeneity of the number of isolates analyzed in the different plots, these were considered positive when at least one isolate was resistant, and otherwise negative. Data were observed at continuous locations occurring within a defined spatial domain (geostatistical data). Therefore, the spatial effect was included via the stochastic partial differential equation approach (SPDE) [1]. The model showed a strong effect of the spatial component, resulting in a high probabiliy of presence of the resistant fungus in areas next to the positive plots. These results suggest that the dissemination of *P. nawae* by ascospores would have a decisive role in the spread of resistant isolates.

Keywords: Spatial, Hierarchical Bayesian models, INLA

References

[1] Lindgren, F., H. Rue, and J. Lindström. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 3(4): 423–498.

[2] Rue, H., S. Martino, and N. Chopin. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1(2): 319–392.

Roadmap for assesing statistical modeling of relative biomass indices

Sep 15th
09:30

A. Fuster-Alonso[1], D. Conesa[2], S. Cerviño[3], M. Cousido-Rocha[3], F. Izquierdo[3], M.G. Pennino[3]

[1] Máster en Bioestadística, Universidad de Valencia; [2] Departamento de Estadística e Investigación Operativa (VaBaR), Universidad de Valencia; [3] Centro Oceanográfico de Vigo, IEO-CSIC

Estimating changes in fish biomass behavior over time is crucial for monitoring species population. However, knowing the totaly of species real biomass in their environment along space-time is not feasible. Thus, statistical modeling appears as a necessary tool to capture the behavior of biomass in space and time by converting several sources of information available. Among them, the most common sources of information are **relative biomass** and **CPUE (catches per unit effort) indices** derived, respectivily, from oceanographic surveys and fishing activity.

The goal of this study is to shed light and help scientists to choose amount statistical methods proposed in the literature for modeling the relative indices of biomass in order to be representative of the real population biomass. For this purpose, we have delevoped de following protocol:

1. **Simulate** a spatio-temporal stock biomass scenario.
2. **Sample** from the simulated biomass using two different aprroaches (i.e., **random sampling** as in a oceanographic surveys and **preferential sampling** as in fishing activities) and get the relative biomass and CPUE indices.
3. Applicate different **statistical models** to obtain the predicted series of relative biomass and CPUE indices; and
4. **Analyze** the predicted series with respect to the simulated biomass.

At point 3) our protocol considers GLMs (Generalized Linear Models), GAMs (Generalized Additive Models), Geostatistical Models and Preferential Models. Inferences on these models is performed within the **frequentist** and **Bayesian** approaches when possible. To compare the results of the different models at point 4) we used error measures, specifically the RMSE (root mean square error) and MAPE (mean absolute percentage error).

The results of this research show that the model that best reflects the behavior of simulated biomass over time is the **geostatistical** one applied to data derived from oceanographic surveys. Moreover, for the CPUE indices derived from fishing activity, the model that best capture the behavior of simulated bimass is **preferencial model**.

Finally, by using the protocol described in this study, we argue that ignoring the underlying **spatial process** in the statistical models can lead to less accurate relative biomass and CPUE indices. In fact, our protocol confirms that we should be careful when modeling the relative biomass and CPUE indices, because a **wrong model** may result in useless or completely **incorrect conclusions**.

Keywords: geostatistics, simulation and statistical modeling

Sep 15th
09:45

Multivariate spatio-temporal models for predicting short-term cancer incidence

Garazi Retegui[1,2], Jaione Etxeberria[1,2], Andrea Riebler[3] and María Dolores Ugarte[1,2]

[1] Department of Statistics, Computer Science and Mathematics, Public University of Navarre (UPNA), Arrosadia Campus, 31006, Pamplona, Spain; [2] Institute for Advanced Materials and Mathematics (INAMAT2), Public University of Navarre (UPNA), Arrosadia Campus, 31006, Pamplona, Spain; [3] Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Alfred Getz' vei 1, 7034, Trondheim, Norway

For a correct allocation of health resources aimed at cancer prevention and control, different indicators such as cancer incidence and mortality rates (or counts) are considered. In most countries, cancer mortality figures are recorded by Statistical Offices and they are available at different levels (national, provincial, municipal or even in census tracts). On the other hand, cancer incidence figures are routinely recorded by national or regional population-based cancer registries. However, these figures are usually available with a delay and therefore, methods that provide short-term predictions are very useful. In this context, an additional drawback is encountered. In large countries, regional cancer registries are responsible for collecting and identifying all cancer cases occurring in a certain domain (state or province for example). Regional registries are usually not established in the same year and therefore, cancer incidence data series between different regions of a country are not harmonised over time. This lack of information (mainly at the beginning of the data series), makes it difficult to use univariate spatio-temporal models. One possibility to solve this problem is to use cancer mortality data as an additional source of information. Therefore, with the aim of predicting cancer incidence in the short-term, we use different multivariate spatio-temporal models which modelled jointly cancer mortality and incidence data. The performance of these multivariate models will be analysed using lung and prostate cancer incidence and mortality data during the period 2001-2015 reported by the 16 regional cancer registries of Germany and provided by the German Centre for Cancer Registry Data, ZfKD.

Keywords: Cancer incidence projection, Multivariate space-time modelling, Predictive accuracy

---

Sesión 4: Supervivencia
15 de septiembre, 10:00 a 11:00

Chair: Juan Carbonell Asíns

Bayesian Survival Analysis of Acute-On-Chronic Liver Failure in Clinically Stable Outpatients with Cirrhosis

Sep 15th
10:00

Pablo Escobar[1], *Carlos Peña*[1], *María Pilar Ballester*[2], *Thomas Tranah*[3], *Debbie Shawcross*[3], *Rajiv Jalan*[4,5], *Juan Carbonell*[1]

[1]Unidad de Bioinformática y Bioestadística. Insituto de Investigación Santiaria (INCLIVA); [2] Digestive Disease Department, Hospital Clínico Universitario de Valencia, Spain; [3]Institute of Liver Studies, Dept of Inflammation Biology, School of Immunology and Microbial Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom; [4] Liver Failure Group, Institute for Liver and Disease Health, University College London, Royal Free Campus; [5] European Foundation for the Study of Chronic Liver Failure (EF Clif) and the European Association for the Study of the Liver-Chronic Liver Failure (EASL-CLIF) Consortium

Corresponding author: jacarbonell@incliva.es

Ammonia level correlates with the severity of hepatic encephalopathy and organ failure and is an independent predictor of complications in patients with acute decompensation or acute-on-chronic liver failure (ACLF). However, its utility as a prognostic biomarker in patients with cirrhosis remains unclear. We hypothesized that hyperammonaemia predisposes to ACLF in outpatients with cirrhosis.

A prospective observational study of clinically stable cirrhotic outpatients followed-up in three tertiary hospitals was performed. Considering there are two types of possible competing events -ACLF and liver transplantation- a univariable competing risk modeling was performed. ACLF is contemplated as the event of interest and liver transplantation as a competing risk. Survival analysis was implemented using a BUGS syntax that can be run with JAGS from the R programming language.

We performed univariable models using two different baseline hazard functions: a Weibull distribution and a mixture of piecewise constant functions, which allows for a more flexible adjustment. Moreover, we tested including a multiplicative frailty term using the individual effect to gathered the heterogeneity present in the data coming from the three different hospitals. In all models analysed, ULN (Upper-Limit Ammonia) appears as a strong risk factor, with a probability of having a regression coefficient greater than 1 of 1 and a relative risk ratio that oscillates between 2.09 and 3.33 for the 95% credible interval in the models analysed. Therefore, this results confirmed the possibility of using ULN as prognosis tool for ACLF in patients suffering from cirrhosis.

Finally, we expect to implement a bayesian variable selection process to select the right combination of covariates, with the idea of implementing in the next months multivariate models that could be a starting point for a future score that could be use as a prognosis tool in health services.

Keywords: Survival Analysis, Bayesian, Competing Risks.

Sep 15th
10:15

A Bayesian spatial illness-death model to assess geographical differences in the risk and incidence of recurrent hip fracture and death.

Fran Llopis-Cardona[1], Carmen Armero[2], Gabriel Sanfélix-Gimeno[1,3]

[1]Health Services Research Area, Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO), Spain; [2]Department of Statistics and Operations Research, Universitat de València, Spain; [3]Network for Research on Chronicity, Primary Care, and Health Promotion (RICAPPS).

Illness-death models are stochastic models, inside the multi-state framework, in which individuals are allowed to move over time between different states regarding illness and death. They allow to study progression through different health conditions by means of transition probabilities. We propose a spatial illness-death model, defined using Cox proportional hazards models, which include multivariate random effects for assessing both correlation between transitions and between spatial units. We apply this model to a cohort study of recurrent hip fracture in older patients. Data come from the PREV2FO cohort, including patients aged 65 years and older who were discharged alive after hospitalization due to a hip fracture during 2008-2015 in the Valencia region, Spain. We assess geographical differences in the incidence of recurrent hip fracture, as well as transition probabilities of refracture, total death and death after refracture. We use a Bayesian approach using the integrated nested Laplace approximation (INLA).

Keywords: Bayesian inference, multi-state models, spatial models

Modelos aditivos y multiplicativos de supervivencia: Un estudio comparado

Sep 15th
10:30

Javier Martín-Pozuelo Lozano[1], Jose Domingo Bermúdez Edo[2]

[1]Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (Fisabio); [2]Departament d'Estadística I Investigació Operativa. Facultat de Ciències Matemàtiques, Universitat de València.

El análisis de supervivencia es una rama de la Estadística empleada para analizar datos procedentes de estudios científicos relativos a tiempos de ocurrencia de uno o varios eventos de interés. En este contexto, el modelo de riesgos proporcionales de Cox [1] es el modelo estadístico semiparamétrico más empleado para trabajar con variables que suponen un único evento de fallo. Sin embargo, este modelo asume una hipótesis muy restrictiva que debe verificarse para validar la aplicación práctica de este, la hipótesis de riesgos proporcionales. Es por ello que Ling y Ying [2] realizaron esfuerzos en obtener una estimación semiparamétrica de la función de riesgo acumulado y los coeficientes del modelo asumiendo una relación aditiva. En este trabajo se implementa la estimación semiparamétrica llevada a cabo por Ling y Ying, así como metodología de simulación de datos de supervivencia de acuerdo con las asunciones de las hipótesis de ambos modelos. Esta implementación permite evaluar y comparar, mediante un análisis de sensibilidad, la robustez de las estimaciones de los coeficientes y las funciones de riesgo acumulado y supervivencia obtenidas por ambos modelos, así como evaluar la importancia del contraste de riesgos proporcionales sobre el modelo de Cox, aplicando ambos modelos a distintos escenarios simulados. A través de este estudio se consiguió analizar la diferencia en las estimaciones obtenidas del modelo de riesgos aditivos y el modelo de Cox, mostrando una mejora en la precisión de la función de supervivencia estimada del modelo aditivo frente al modelo de riesgos proporcionales.

Keywords: Análisis de supervivencia, análisis de sensibilidad, modelización estadística

References

[1] Cox,D.R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 34(2), 187-220.

[2] Lin,D.Y., and Ying,Z. (1994). Semiparametric analysis of the additive risk model. *Oxford University Press on behalf of Biometrika Trust.*

Sep 15th
10:45

Asociación entre el cáncer de cuello uterino y las medidas antropométricas y la actividad física

Jon Aritz Panera Carracedo
[1]Programa de Recerca en Epidemiologia del Càncer (PREC)

Introducción. El cáncer de cuello uterino figura tercero en la lista de enfermedades cancerosas, tanto si se habla de incidencia como de mortalidad. El Virus del Papiloma Humano es causa necesaria del cáncer de cuello uterino. Se han estudiado varios factores de riesgo: tabaquismo, uso prolongado de anticonceptivos orales, alta paridad o otras infecciones de transmisión sexual como la clamidia, el herpes o el VIH. En cuanto al índice de masa corporal (BMI; por su nombre en inglés) y la actividad física (AF) existen asociaciones aunque éstas son inconsistentes.

Objetivos. El objetivo de este trabajo es estudiar la asociación entre las medidas antropométricas y la actividad física, y el riesgo de desarrollar un cáncer de cuello uterino o uno de sus precursores. Además, se desarrollará un algoritmo para aplicar una bondad de ajuste para datos emparejados en un modelo de regresión logística condicional.

Métodos. A partir de las bases de datos del estudio EPIC (European Prospective Investigation into Cancer and Nutrition), se realizarán modelos de riesgo proporcional de Cox para estimar *hazard ratios* y modelos de regresión logística condicional para estimar *odds ratios*. Además, se validarán ambos tipos de modelos, en concreto, el de regresión logística condicional mediante una bondad de ajuste específico para datos emparejados.

Resultados. El riesgo instantáneo de padecer una lesión precancerosa de cuello uterino es un 49.2% más alto en mujeres con un BMI de peso bajo, en comparación mujeres con índice saludable y mismas características en las demás variables. Las mujeres con obesidad, estadísticamente, representan un factor protector en el precáncer. Este resultado está sujeto a un sesgo por la falta de cribado de estas mujeres. El riesgo instantáneo de que una mujer con una actividad física activa padezca un cáncer invasor es un 24.4% menor en comparación con una mujer inactiva y mismas características en las demás variables.

Conclusiones. El algoritmo de la bondad de ajuste para datos emparejados sirve para identificar estratos e individuos influyentes o mal ajustados. Un estilo de vida con una actividad física activa y un índice de masa corporal saludable es recomendable.

Sesión 5: Machine learning
15 de septiembre, 11:30 a 12:30

Chair: Harold Antonio Hernández Roig

Machine-learning use of risk prediction models to triage the severity level of COVID-19 patients entering the emergency care system

Sep 15th
11:30

Goizalde Badiola-Zabala[1], Jose Manuel Lopez-Guede[1,2], Manuel Graña[1,3]

[1]Computational Intelligence Group, Basque Country University (UPV/EHU); [2]Dept. of Systems and and Automatic Control, Faculty of Engineering of Vitoria, Basque Country University (UPV/EHU), Nieves Cano 12, 01006, Vitoria-Gasteiz, Spain; [3]Dept. of Computer Science and Artificial Intelligence, Faculty of Informatics, Basque Country University (UPV/EHU), Paseo Manuel de Lardizabal 1, 20018, Donostia-San Sebastian, Spain

The upsurge in coronavirus cases has led to a large influx of patients to hospitals around the world. This new and challenging situation has intensified the emergence of clinical decision-making systems. Such technologies have been used to alleviate the unbearable strain on healthcare systems.

A concise project has been carried out with several advances achieved so far in the use of ML technologies in pandemic management, as well as its application in the use case presented, carrying out a thorough methodology to carry out the project, trying to answer the research questions posed, counting on the data set given by a local hospital that will be processed for the developments, and attempting to mention several improvements that will be applied to the data set.

In an attempt to address this challenging healthcare situation, we attempt to make a contribution to this effort with an immersive study of a real data set of COVID-19 patients from a local hospital. In this study, we approach the problem as a triage prediction problem, formulated as a multiclass classification problem, with special attention to the age normalization of physiological variables, i.e., heart rate and blood pressure. We report experimental results obtained on a data sample consisting of COVID-19 patients attended in a local hospital. To this end, we endeavored to emulate the triage decisions of the physicians registered in a dataset containing the measurements of the demographic variables and physiological variables (vital signs) and the triage decision. We obtained results that provide incentive toward the development of a real-life application of data balancing and classification into the triage prediction that clinicians designate to critically ill patients.

Keywords: COVID-19, Modeling, Machine Learning

Sep 15th
11:45

Exploring statistical methods for classifying individuals in extreme aging groups

Armand González-Escalante[1], Blanca Rodríguez-Fernández[1], Irene Cumplido-Mayoral[1], Juan Domingo Gispert[1], Marta Crous-Bou[1],Natalia Vilor-Tejedor[1], Marc Suárez-Calvet[1]

[1]Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation.

Background: Aging is the most important risk factor for Alzheimer's disease (AD) and other dementias. A better understanding of how individual biological and brain ages can provide important information upon which to base strategies for new therapies. The purpose of this study was to explore statistical metrics to classify individuals in extreme aging groups according to biomarkers that are related to individual variability in rate of aging.

Methods: This classification resulted in the obtention of 340 middle-aged cognitively unimpaired participants at risk of AD from the Alzheimer's and Families (ALFA) cohort with an extreme aging profile. For the classification, we used telomere length (TL) and magnetic resonance imaging (MRI) brain features for computing two aging metrics. Telomere length (TL) was determined by qPCR from DNA extracted from peripheral blood leukocytes. The delta-age from TL data was calculated as the residuals from regressing chronological age on TL z-scores in women and men. Moreover, we used brain features to compute a BrainAge metric. The model of healthy brain aging was first trained with the chronological age and pre-processed structural MRI data of a training sample using a gradient boosting algorithm for capturing the multidimensional aging patterns throughout the whole brain. Then, the delta-age for the BrainAge metric was calculated on the testing sample as the difference between the estimated and chronological age using the already trained model. Aged and rejuvenated aging groups were defined as those 85 more extreme individuals of the computed metrics, roughly corresponding to the 10th and 90th percentile.

Results: Significant differences were found between aging groups and between sexes for both metrics, showing the classification ability of both estimated aging metrics (p-values < 0.001). Most of the other demographic variables tested didn't show significant differences between the groups (p-values > 0.05), with just BMI and Years of education, being different between the men extreme groups defined by the BrainAge metric.

Conclusions: The comparisons performed in this preliminary study confirmed the classification ability of the calculated age metrics between the groups. Future analyses will focus on the analysis of circulating blood factors that differ between groups, integrating proteomics and metabolomics data to unravel the biological mechanisms associated with variability in the rate of aging.

Keywords: Aging, Bioinformatics, Classifying methods

Development and validation of prognostic models for hospitalization in the Basque Country: Analyzing the variability of non-deterministic algorithms

Sep 15th
12:00

Alexander Olza Rodríguez[1], Eduardo Millán Ortuondo[2,3], María Xosé Rodríguez-Álvarez[4,5]

[1]Basque Center for Applied Mathematics (BCAM), Bilbao, Spain; [2]Osakidetza Basque Health Service, General Directorate for Healthcare; [3] Kronikgune, Institute for Health Services Research [4] CINBIO, Universidade de Vigo, Dept. of Statistics and OR, 36310 Vigo, Spain; [5] CitMAGA, Center for Mathematical Research and Technology of Galicia, Spain.

Introduction: We predict the probability of unplanned hospitalization in the Basque Country and identify risk groups using several techniques. When dealing with non-deterministic algorithms, the comparison of a single model per technique is not enough to choose the best approach. Thus, we evaluate the median performance and variability of three families of models to be able to make an informed decision.

Methods: We conduct 40 experiments per family of models - Random Forest (RF), Gradient Boosting Decision Trees (GBDT) and Multilayer Perceptrons (MLP) - and compare them to Logistic Regression (LR). Except for LR, the other methods are non-deterministic. Hyperparameter tuning and undersampling also add randomness to the modelling. Different realizations of the entire process may thus lead to diverse rankings of the techniques. Controlling all sources of randomness, we study the variability of the three non-deterministic algorithms. We combine Isotonic Regression and PCHIP polynomial smoothing to achieve calibration.

Results: The best-performing technique is MLP, followed by GBDT, LR and RF. MLPs also have the lowest variability, around an order of magnitude less than RF. Median AUC ranges from 0.789 to 0.802. The median Average Precision (AP) is between 0.237 and 0.257 (coin-toss model: AP = 0.0567). Median Recall @20k ranges from 0.076 to 0.080. As for Positive Predictive Value @20k, the median values are between 0.485 and 0.511%. Hence, 50% of the times, 51.1% of the patients shortlisted by MLPs will indeed be admitted the next year. This percentage drops to 48.3% with RF. There is some overlap between the algorithms. For instance, GBDT performs better than LR more than 75% of the time, but not always.

Conclusions: All models have good global discrimination, with AUCs around 0.80 and APs around 0.25. Identifying the 20000 highest-risk patients is much more difficult, resulting in lists with low recall, because we impose a number of positives that is much lower than the admissions per year. However, the positive predictive value (around 50%) indicates that, at the reviewing stage by primary care physicians, the suggested lists will be very useful for clinical practice. Regarding the evaluation of different techniques, we do not see large differences in the metrics. The only family that is consistently superior to LR is MLP, showing a very reliable performance with the lowest variability. Our predictors are highly processed binary variables, leaving little room for machine learning techniques to excel with this specific kind of data-set.

Keywords: Hospitalization, non-deterministic algorithms, predictive models

Modelos de árboles de regresión y categorización aditivos bayesianos: un encuentro entre dos culturas

Sep 15th
12:15

Alfonso Picó[1], Carmen Armero[1], Gianni Gallello[2]

[1]Departamento de Estadística e Investigación operativa, Universitat de València (España);
[2]Departamento de Prehistoria, Arqueología e Historia Antigua, Universitat de València (España)

Desde finales de los años 70 la comunidad estadística ha estado partida en lo que posteriormente algunos autores denominaron "dos culturas". Por un lado, aquellos que especifican un modelo estocástico subyacente que hubiera generado teóricamente los datos de la muestra, modeladores de datos, fijan el énfasis de sus investigaciones en obtener estimaciones que favorecen la inferencia. Por otro lado, estadísticos especialmente centrados en problemas aplicados y cuyo fin último era conseguir una gran precisión en sus predicciones. Estos últimos parecen haber confirmado su independencia con la disciplina de la data science y han hecho del llamado machine learning su paladín.

En el presente trabajo expondremos un caso único que aúna la fortaleza del modelado y la flexibilidad del machine learning gracias al marco de trabajo bayesiano por medio de los árboles de regresión aditivos bayesianos (BART), una familia de algoritmos con un punto de partida común: incorporar la incertidumbre y su cuantificación en la construcción de los árboles de regresión que formarán el ensamblado final.

Pondremos a prueba diferentes algoritmos de predicción de uso habitual frente a un modelo BART en una tarea de clasificación de tipos de huesos. Se explorarán las posibilidades de futuro en la aplicación de los BART y modelos bayesianos semejantes no solo en el campo arqueológico o forense, motivo principal de este trabajo, sino en su posible transferencia al mundo biomédico y la industria.

Keywords: BART, machine learning, bayesiano

Sesión de pósteres
15 de septiembre, 12:30 a 13:30

Modelos de Segmentación Aplicados a la Caracterización del Sector Vacuno Cárnico Español

Sep 15th
12:30

M. Anciones-Polo[1], P. Vicente-Galindo[2], P. Galindo-Villardón[3]

[1]Departmento de Estadística, Universidad de Salamanca; [2]Departamento de Estadística, Universidad de Salamanca; [3]Departamento de Estadística, Universidad de Salamanca

La ganadería, en su totalidad, es una actividad con una reseñable importancia dentro de nuestra sociedad, tanto a niveles económicos, como sociales y ambientales. Esta investigación tendrá como marco específico de estudio del sector vacuno cárnico español, el cual tiene un peso considerable en el medio rural, desempeñando, desde su origen hasta la actualidad, un papel clave en la alimentación, supervivencia y conservación de estas poblaciones, dado su estrecho vínculo con el entorno. Además, se pretende generar conocimiento respecto al análisis de datos multivariantes como herramienta significativa atribuible a este sector, así como elaborar un mapa de vacuno cárnico en el ámbito nacional, con la pretensión de poder plasmar su importancia social y económica, exponiendo también su papel en la conservación de parte de los ecosistemas españoles y en la lucha contra el cambio climático.

La razón de este estudio radica en una mayor concienciación sobre las bondades y debilidades del sector vacuno, mediante la implantación de métodos estadísticos multivariantes que resulten adecuados según la naturaleza de la base de datos a estudio. Utilizaremos el CHAID Y EL TAID como modelos de segmentación, los cuales nos permitirán construir un perfil más preciso de las explotaciones que conforman el mapa nacional, agrupar para conocer subgrupos muestrales y obtener mejores pronósticos sobre el comportamiento de los grandes grupos de datos obtenidos.

Keywords: Ganadería, Sector Cárnico, Vacuno, Técnicas Multivariantes, CHAID, TAID

References

[**1**] Castro-Lopez C., Vicente-Galindo P., Galindo-Villardón P., Borrego-Hernández O. (2005). TAID-LCA: Segmentation Algorithm Based on Ternary Trees. *Mathematics* (2022), 10(4), 560.

[**2**] Djordjevic, D.; Cockalo, D.; Bogetic, S.; Bakator, M. Predicting Entrepreneurial Intentions among the Youth in Serbia with a Classification Decision Tree Model with the QUEST Algorithm. *Mathematics* (2021), 9, 1487.

[**3**] Avila, C.A. Una Alternativa al Análisis de Segmentación Basada en el Análisis de Hipótesis de Independencia Condicionada. Ph.D. Thesis, Universidad de Salamanca, Salamanca, Spain, 1996.

[**4**] Kass, G. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *J. Appl. Stat.* (1980), 29, 127–199.

Aplicación de técnicas estadísticas multivariantes en el análisis, estudio y optimización de los indicadores de lesionabilidad y rendimiento físico en jugadores de fútbol profesional

Sep 15th
12:35

E. Benéitez-Andrés[1], M. Sánchez-Barba[1], M. Sánchez[2]

[1]Departamento de Estadística, Facultad de Medicina, Universidad de Salamanca; [2]Departamento de Ciencias de la Actividad Física y del Deporte, Facultad de Educación, Universidad Pontificia de Salamanca

Nuestros hallazgos, basados en la aplicación de técnicas multivariantes, de reducción de dimensionalidad y clasificación automática, nos permiten aportar herramientas optimizadoras respecto a la mejora respecto al rendimiento físico e incidencia lesional en jugadores de fútbol profesional.

A partir del registro de la carga externa por sesión, microciclo y mesociclo podemos estimar y ajustar, de forma dinámica, la condición y rendimiento físico del jugador, acrecentando la eficiencia del proceso de entrenamiento en este, así como, atenuar los efectos negativos que la actividad deportiva de alto rendimiento tiene sobre el individuo, es decir, reduciendo el tiempo de lesión y el número de lesiones que este sufre (Owoeye, O., 2020[1]). Mediante sistemas de geoposicionamiento se registraron diversas variables de carga externa como son la distancia total, distancia a sprint, distancia a alta (DAV), media y baja velocidad y el número de aceleraciones (ACC), todas ellas expresadas en función del tiempo de participación, además, se registraron los índices de esfuerzo percibido, bienestar diario y se tuvo en cuenta otras variables contextuales como la condición del equipo (local-visitante) y el rol del jugador (titular-suplente) en el día de partido (MD).

Los datos se analizaron según la perspectiva del día de entrenamiento, el tipo de tarea, la carga acumulada por partido, el efecto temporal y acumulativo de la misma. Para apoyarnos en la explicación de los resultados obtenidos se emplearon diversas técnicas de representación gráfica basadas en la reducción de dimensionalidad, estas permiten clasificar y discriminar variables e individuos-observaciones, entre otras HJ-Biplot, Clúster K-Means, Clara y PAM.

Los resultados mostraron como la sesión MD+1 y MD-1 son las que presentan menor capacidad de discriminación respecto a las variables contextuales registradas, careciendo de importancia para el índice de lesionabilidad, las sesiones MD-4, -3 y -2 son las de mayor índice lesional, las variables contextuales jugar local y ser titular aumentan el rendimiento observado en las variables de carga externa ACC y DAV.

Keywords: Análisis multivariante, rendimiento deportivo, lesión

References

[**1**] Owoeye, O., VanderWey, M. J., and Pike, I. (2020). Reducing injuries in soccer (football): an umbrella review of best evidence across the epidemiological framework for prevention. *Sports medicine-open*, 6(1), 1-8.

Estudio sobre el microbioma intestinal, y la calidad de vida y el neurodesarrollo en adolescentes

Sep 15th 12:40

R. Beneyto[1]; B. Sarzo B[1,2,3]; MA. Martinez-Beneito MA[4]; MJ. Lopez-Espinosa [1,3,5,6]

[1]Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana, FISABIO-Salud Pública, Valencia, España; [2]Unidad de investigación mixta en Epidemiología y Salud Ambiental, FISABIO-Universidad Jaume I-Universidad de Valencia, Valencia, España; [3]Departamento de Microbiología y Ecología, Universidad de Valencia, Valencia, España; [4]Departamento de Estadística e Investigación Operativa, Valencia, España; [5]Consorcio Español para la Investigación en Epidemiología y Salud Pública (CIBERESP), Madrid, España; [6]Facultad de Enfermería y Podología, Universidad de Valencia, Valencia, España.

Introducción. El intestino grueso humano se encuentra habitado por un gran número de microorganismos, y su composición varía entre individuos según diversos factores. Durante los últimos años ha habido un creciente interés por estudiar la relación entre los microorganismos del intestino grueso, y funciones cerebrales y salud mental, relación que se conoce como eje bidireccional microbiota-intestino-cerebro. La mayoría de los estudios realizados hasta la fecha se han centrado en población adulta, siendo la adolescencia una etapa de la vida poco estudiada.

Objetivo. Estudiar la posible relación entre la microbiota intestinal y la calidad de vida y el neurodesarrollo en población adolescente.

Bases de datos y métodos de análisis. La población de estudio estaba formada por 192 adolescentes de Valencia (edad = 14-16 años), de los que se obtuvo información sobre microbiota intestinal de muestras fecales en forma de conteos absolutos para cada categoría taxonómica e índices de diversidad α (CHAO1 y Shannon). Además, se calculó la diversidad β para dividir a los individuos en grupos según su composición. Por otro lado, se obtuvo información sobre calidad de vida y neurodesarrollo a partir de diferentes cuestionarios y pruebas. Para estudiar la posible asociación entre microbiota y los resultados de los cuestionarios y pruebas se utilizó regresión lineal simple (diversidad α), ANOVA (diversidad β) y métodos de selección de variables (composición bacteriana a nivel taxonómico de familia).

Resultados. Se encontró una sutil relación negativa entre uno de los índices de diversidad α (CHAO1) y uno de los indicadores de calidad de vida (indicador de autonomía y relaciones parentales, p-valor = 0,02). En los análisis para la diversidad β se observaron diferencias sutiles entre grupos de individuos para tres indicadores de calidad de vida (autonomía y relaciones parentales [p-valor = 0,045], bienestar emocional [p-valor = 0,025], y entorno escolar [p-valor = 0,033]). En el caso de la composición bacteriana, los métodos de selección de variables arrojaron pocos resultados significativos, encontrando solamente una posible relación positiva entre una familia bacteriana y uno de los indicadores de calidad de vida (indicador global).

Conclusiones. Las asociaciones encontradas entre la microbiota intestinal y los indicadores de calidad de vida son sutiles. En futuros estudios se aumentará el tamaño muestral y se realizarán ajustes por variables sociodemográficas y antropométricas para optimizar los análisis estadísticos.

Keywords: Microbiota Intestinal, Salud Pública, Estadística

Sep 15th
12:45

Técnicas CHAID Aplicadas a la Caracterización del Sector Vacuno Cárnico Español.

Anciones-Polo, M.[1], Vicente-Galindo, P.[1], Galindo-Villardón, P.[1]
[1]Departmento de Estadística, Universidad de Salamanca

La ganadería, en su totalidad, es una actividad con una reseñable importancia dentro de nuestra sociedad, tanto a niveles económicos, como sociales y ambientales. Esta investigación tendrá como marco específico de estudio del sector vacuno cárnico español, el cual tiene un peso considerable en el medio rural, desempeñando, desde su origen hasta la actualidad, un papel clave en la alimentación, supervivencia y conservación de estas poblaciones, dado su estrecho vínculo con el entorno. Además, se pretende generar conocimiento respecto al análisis de datos multivariantes como herramienta significativa atribuible a este sector, así como elaborar un mapa de vacuno cárnico en el ámbito nacional, con la pretensión de poder plasmar su importancia social y económica, exponiendo también su papel en la conservación de parte de los ecosistemas españoles y en la lucha contra el cambio climático.

La razón de este estudio radica en una mayor concienciación sobre las bondades y debilidades del sector vacuno, mediante la implantación de métodos estadísticos multivariantes que resulten adecuados según la naturaleza de la base de datos a estudio. Utilizaremos el CHAID Y EL TAID como modelos de segmentación, los cuales nos permitirán construir un perfil más preciso de las explotaciones que conforman el mapa nacional, agrupar para conocer subgrupos muestrales y obtener mejores pronósticos sobre el comportamiento de los grandes grupos de datos obtenidos.

Keywords: Ganadería, Sector Cárnico, Vacuno, Técnicas Multivariantes, CHAID, TAID

References

[**1**] Castro-Lopez C., Vicente-Galindo P., Galindo-Villardón P., Borrego-Hernández O. (2005). TAID-LCA: Segmentation Algorithm Based on Ternary Trees. *Mathematics (2022), 10*(4), 560.

[**2**] Djordjevic, D.; Cockalo, D.; Bogetic, S.; Bakator, M. Predicting Entrepreneurial Intentions among the Youth in Serbia with a Classification Decision Tree Model with the QUEST Algorithm. *Mathematics (2021), 9*, 1487.

[**3**] Avila, C.A. Una Alternativa al Análisis de Segmentación Basada en el Análisis de Hipótesis de Independencia Condicionada. Ph.D. Thesis, Universidad de Salamanca, Salamanca, Spain, 1996.

[**4**] Kass, G. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *J. Appl. Stat. (1980), 29*, 127–199

APLICACIÓN DEL BOOTSTRAP

Sep 15th
12:50

Gresky Oscar Gutiérrez [1]

[1]Departmento de Estadística, Universidad de Salamanca

Sabemos que la estadística es la ciencia que aprende desde la experiencia, experiencia que se va acumulando poco a poco en el tiempo.

La teoría estadística proporciona métodos óptimos para encontrar una señal real en un fondo ruidoso y proporciona de igual forma controles estrictos contra la sobre interpretación de patrones aleatorios.

Los métodos estadísticos basados en muestreo con reposición se engloban dentro de las estadísticas no paramétricas; las cuales se basan en carecer de distribuciones específicas, y se utilizan cuando las condiciones de las pruebas paramétricas no se cumplen o cuando se quiere hacer inferencia estadística sobre un parámetro distinto a la media.

El propósito de este póster es describir y aplicar uno de los métodos de re-muestreo más utilizado: EL BOOTSTRAP.

En la teoría el escenario ideal para realizar inferencia sobre una población es disponer de una gran cantidad de muestras de dicha población. En la práctica no suele ser posible acceder a múltiples muestras. Si sólo se dispone de una muestra, y esta es representativa de la población cabe esperar que los valores en la muestra aparezcan aproximadamente en la misma frecuencia que en la población.

El método bootstrap [1] se basa en generar nuevas pseudo-muestras, del mismo tamaño que la muestra original mediante el muestro con reemplazo de los datos disponibles.

Si la muestra original es representativa de la población, la distribución del estadístico calculada a partir de las pseudo-muestras (distribución bootstrap) se aproxima a la distribución muestral que se obtendría si se pudiera acceder a la población para generar nuevas muestras.

Keywords: Bootstrap, Jackknife

References

[1] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*,7, 1-26.

[2] Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, jackknife, and cross-validation. *American Statistician*, 37, 36-48.

[3] Efron, B. and Tibshirani, R. J. (1993). An introduction to the bootstrap. New York:Chapman & Hall.

Clinical prediction rules for adverse outcomes in patients with SARS COV-2 infection by the omicron variant

Sep 15th
12:55

Lander Rodríguez[1], Irantzu Barrio [1,2], Ane Villanueva [3,4], Jose María Quintana-Lopez[3,4]

[1]Applied Statistics Group, Basque Center for Applied Mathematics; [2]Department of Mathematics, University of the Basque Country UPV/EHU; [3]Galdakao-Usansolo University Hospital, Respiratory Unit, Osakidetza Basque Health Service; [4]Kronikgune Institute for Health Services Research, Barakaldo, Spain

The global pandemic of COVID-19 has caused millions of deaths throughout the world, but many of its aspects remain unknown. In consequence, factor identification of negative outcomes such as death, adverse evolution (ICU or death) and hospitalization is necessary. The present study is a retrospective cohort study of patients with SARS-CoV-2 infection from March 1st 2020 to January 9th 2022. Data collected for this study included sociodemographic data, baseline comorbidities and treatments. The period from March 1st 2020 to December 13th 2021 was considered as a sample for model development (Derivation Data Set), while the period from December 14th 2021 to January 9th 2022 (Omicron) was used as a validation sample. The Derivation Data Set was randomly divided in equal halves. One half (50%) was used for variable selection and estimation of parameters of the prediction model (train) and the other half (50%) was used for internal validation (test). Multivariable logistic regression models were developed with Lasso logistic regression in the train subsample for parameter estimation and variable selection. In the final models, only factors with $p < 0.01$ were retained. Predictive risk scores for each of the outcomes were developed, by first assigning a weight to each risk predictor variable in relation to the estimated β parameters based on the lasso logistic regression model derived in the train subsample. Then risk weights of all the patient's predictor variables were added up. The discrimination ability of both models and scores was measured by the area under the ROC curve (AUC) The AUCs obtained in the omicron sample ranged from 0.94 (death) to 0.79 (hospitalization). From the previous scores, four risk scales with high predictive capacity were proposed for adverse outcomes and hospital admission. Finally, a shiny application that incorporates these models so that they can be used in clinical practice was developed.

Keywords: COVID-19, Health Care

The food traffic light that gives a green light to ultra-processed foods: visual data mining

Sep 15th
13:00

Carmen Romero Ferreiro[1,2,3], Pilar Cancelas Navia[1,2], David Lora Pablos[1,2,4]

[1] Scientific Support Unit (i+12), Hospital Universitario 12 de Octubre, Madrid, Spain; [2] Spanish Clinical Research Network (SCReN), Madrid, Spain; [3] Facultad de Ciencias de la Salud, Universidad Francisco de Vitoria, Pozuelo de Alarcón, Madrid, Spain; [4] Facultad de Estadística, Universidad Complutense de Madrid (UCM), Madrid, Spain

Simple correspondence analysis is a descriptive method of analysis that graphically represents tables of data [1]. The input data for correspondence analysis forms a contingency table showing the relationship of two variables. When one of the variables has three categories, a triangular coordinate system (ternary diagram) is used to graphically represent this relationship. This type of analysis allows to visualize the relationship between two variables and can be very useful in health sciences. In this study, simple correspondence analysis was used to describe the relationship between two food nutritional value ranking systems, Nutri-Score and the NOVA classification. Nutri-Score [2] is a front-of-pack labelling that classifies foods according to their nutritional quality using five letters (A, B, C, D and E), each associated with a colour, but does not consider other dimensions such as the degree of food processing. The NOVA classification [3] divides foods according to their degree of processing into 4 groups. The Open Food Facts database was used to obtain all products currently marketed in Spain with the Nutri-Score and NOVA classification (n=9931). The relationship between the two categorical variables that classify foods (NOVA and Nutri-Score) was represented by a ternary diagram. Specifically, the five letters of the Nutri-Score (represented by dots) and the percentage of ultra-processed foods according to the NOVA classification (represented on each of the three axes) compose the diagram. Ultra-processed foods (corresponding to the NOVA 4 group) were found in all Nutri-Score categories, ranging from 26.08% in nutritional category A, 51.48% in category B, 59.09% in category C, 67.39% in category D to up to 83.69% in nutritional category E. This can lead to the mistaken assumption that unhealthy foods, such as ultra-processed foods, are "less bad" by consumers. These findings also highlight the need for improve public health tools, such as accompanying the Nutri-Score labelling with complementary labelling indicating the level of processing.

Keywords: Correspondence analysis; Ultra-processed foods; Food labelling

References

[1] Greenacre M.J. (2008) La Práctica del análisis de correspondencias. *Fundación BBVA Baquero*

[2] Galan P. and Babio N. (2019) Nutri-Score: el logotipo frontal de información nutricional útil para la salud pública de España que se apoya sobre bases científicas. *Nutrición Hospitalaria*, 10

[3] Monteiro C.A., Cannon G., Levy R., Moubarac J-C., Jaime P., Martins A.P., et al. (2016) NOVA. The star shines bright. *World Nutrition*, 7(1-3), 28-38

Sesión 6: Biología y ecología
15 de septiembre, 15:15 a 16:15

Chair: Martina Cendoya Martínez

Sep 15th
15:15

Analysis of longitudinal Microbiota data using a Dirichlet Autoregressive Model

I.Creus-Martí[1,2], A. Moya[2,3,4], F.J. Santonja[1]

[1]Department of Statistics, University of Valencia; [2]Institut for Integrative Systems Biology (I2SysBio), University of Valencia; [3]The Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO); [4]The Biomedical Research Center Network for Epidemiology and Public Health (CIBERESP)

Analysing the microbiota dynamics is important due to the link between microbiota and health. We present a Dirichlet autoregressive model that takes into account the compositional characteristics of the microbiome datasets. The model is estimated with maximum likelihood using a strategy to speeds up the estimation. This model allows extracting information about the bacterial behaviour through the interpretation of its parameters and also shows the relationship between microbiome variability and host health status.

Keywords: Dirichlet Autoregressive Model, Microbiota, Compositional Data

References

[**1**] Creus-Martí, I., Moya, A., Santonja, F.J (2021). A Dirichlet Autoregressive Model for the Analysis of Microbiota Time-Series Data. *Complexity*. `https://doi.org/10.1155/2021/9951817`.

Records tests and applications to climate change

Sep 15th 15:30

Jorge Castillo-Mateo[1], Ana C. Cebrián[1], Jesús Asín[1]

[1]Department of Statistical Methods, University of Zaragoza

An observation X_i in a time series (X_t) is called an upper (lower) record if it is greater (smaller) than all previous observations in the series, i.e., $X_i > \max_{t<i}\{X_t\}$ ($X_i < \min_{t<i}\{X_t\}$). The study of records is of interest in different fields, but we focus on record-breaking events in climatology and their connection with climate change. Under the setup where (X_t) is a series of i.i.d. random variables, the binary variables (I_t) that indicate the record occurrence at time t are independent and have a Bernoulli distribution with probability $p_t = 1/t$. This property was first used by Foster and Stuart [1] to develop statistics based on the number of records whose distribution does not depend on the distribution of X_t. The underlying idea to build distribution-free tests is to use the distribution of the record occurrence in an i.i.d. series, and to study if the observed records are compatible with that behavior. In this presentation we review and propose non-parametric trend tests [1,2,3], change-point detection tests [4] and graphical tools [3] based on the record occurrence. We also show the R package **RecordTest** [5,6] available from CRAN which implements all of these tools. Finally, we apply the functions of the package **RecordTest** to temperature series located around the Iberian Peninsula.

Keywords: daily temperature; distribution-free tests; record-breaking events

References

[1] Foster F.G., and Stuart A. (1954). Distribution-free tests in time-series based on the breaking of records. *Journal of the Royal Statistical Society Series B (Methodological)*, 16(1), 1–22.

[2] Diersen J., and Trenkler G. (1996). Records tests for trend in location. *Statistics*, 28(1), 1–12.

[3] Cebrián A.C., Castillo-Mateo J., and Asín J. (2022). Record tests to detect non stationarity in the tails with an application to climate change. *Stochastic Environmental Research and Risk Assessment*, 36(2), 313–330.

[4] Castillo-Mateo J. (2022+). Distribution-free changepoint detection tests based on the breaking of records. *Environmental and Ecological Statistics*, (Accepted). `arXiv:2105.08186 [stat.ME]`.

[5] Castillo-Mateo J., Cebrián A.C., and Asín J. (2022+). **RecordTest**: An R package to analyse non-stationarity in the extremes based on record-breaking events. (Under review.)

[6] Castillo-Mateo J. (2021). **RecordTest**: Inference tools in time series based on record statistics. *R package version 2.1.0*, `http://CRAN.R-project.org/package=RecordTest`.

Evaluación de la idoneidad climática de la cuenca Mediterránea para el desarrollo de la mancha negra de los cítricos, causada por *Phyllosticta citricarpa*

Sep 15th
15:45

Anaïs Galvañ[1], Naima Boughalleb-M'Hamdi[2], Najwa Benfradj[2], Sabrine Mannai[2], Elena Lázaro[1], Antonio Vicent[1]

[1] Institut Valencià d'Investigacions Agràries (IVIA), Centre de Protecció Vegetal i Biotecnologia, 46113 Moncada, Valencia, Spain. [2] Institut Supérieur Agronomique de Chott Mariem, LR21AGR05, University of Sousse, Department of Biological Sciences and Plant Protection, Sousse, 4042, Tunisia.

La mancha negra de los cítricos o '*Citrus black spot*' (CBS) causada por el hongo *Phyllosticta citricarpa*, es la principal enfermedad fúngica de los cítricos a nivel mundial. Estudios previos indicaron que esta enfermedad no es capaz de desarrollarse en condiciones de clima Mediterráneo. No obstante, en 2019 se describió por primera vez la presencia de CBS en Túnez. El objetivo del presente estudio es evaluar la idoneidad climática de la cuenca Mediterránea para el desarrollo del CBS, empleando un modelo genérico de infección para simular las infecciones potenciales por ascosporas y picnidiosporas de *P. citricarpa* y un modelo de grados-día para predecir el inicio de la liberación de ascosporas. Para la caracterización climática de las zonas citrícolas de la cuenca Mediterránea, así como de otras localizaciones afectadas por CBS, se emplearon datos de alta resolución espacial (9 km) extraídos de ERA5-Land. En el modelo genérico de infección se consideraron dos escenarios de simulación, en el primero los parámetros del modelo se estimaron con valores procedentes de bibliografía y en el segundo mediante un proceso inferencial bayesiano. Los resultados del modelo genérico de infección indicaron que las infecciones por ascosporas y picnidiosporas se concentraban principalmente en los meses de otoño, y también en primavera para las picnidiosporas. El modelo grados-día predijo el inicio de la liberación de ascosporas a finales de primavera, aunque al tratarse de un modelo empírico su extrapolación a la cuenca Mediterránea es cuestionable. A diferencia de estudios previos, el modelo genérico de infección simuló un porcentaje de horas favorables para las infecciones por picnidiosporas superior al de ascosporas. Los valores simulados en Túnez y algunas zonas afectadas por CBS fueron, en general, similares a los de las regiones citrícolas de Europa y el Magreb, zonas en las que no se ha descrito la enfermedad. Estos resultados confirman la idoneidad climática de la cuenca Mediterránea para el desarrollo del CBS.

Keywords: CBS, modelo genérico de infección, cuenca Mediterránea, ERA5-Land

Forecast of temperature-attributable mortality at lead times of up to 15 days for a very large ensemble of European regions

Sep 15th 16:00

Marcos Quijal-Zamorano[1,2]; Desislava Petrova[1]; Hicham Achebak[1]; Èrica Martínez-Solanas[1]; Jean-Marie Robine[3,4]; François R. Herrmann[5]; Xavier Rodó[1,6]; Joan Ballester[1]

[1]ISGlobal, Barcelona, Spain; [2]Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain; [3]Institut National de la Santé et de la Recherche Médicale (INSERM), Montpellier, France; [4]École Pratique des Hautes Études, Paris, France; [5]Division of Geriatrics, Department of Rehabilitation and Geriatrics, Geneva University Hospitals and University of Geneva, Thônex, Switzerland;[6]ICREA, Barcelona, Spain

Implementing adequate health preventing measures is essential for public health decision making, particularly in the current context of rising temperatures. Most of the early warning systems are only based on climate data, and in very few cases they truly model the actual impact of the climate phenomena. Here we establish, for the first-time, the theoretical basis for the development of operational heat-health early warning systems that combine climate and health data. We studied the predictability of Temperature Attributable Mortality (TAM) at lead times of up to 15 days for a very large ensemble of European regions. To achieve this goal, we analysed daily counts of all-cause mortality for the period 1998-2012 in 147 NUTS2 regions in 16 European countries, representing more than 400 million people, and daily high-resolution weather forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF). We applied epidemiological models for the fitting of the temperature-mortality relationship in each of the regions, accounting for the different vulnerabilities and socio-demographic characteristics existing in Europe. We compared the predictive skill of the temperature and health forecasts on seasons and days with higher mortality risk. Our results indicate that climate predictability could in fact be the major limiting factor in the design of new early warning systems that also incorporate health information. These results open a new avenue for the transformation of operational weather and subseasonal-to-seasonal climate forecasts into early warning systems of heat, but also a range of other climate-driven human health risks.

Sep 15th
16:15

Biplot logistico asociado al análisis de la redundancia para datos de respuesta binaria

Laura Vicente-Gonzalez[1], Jose Luis[1]

[1]Department of Statistics, Universidad de Salamanca.

El análisis de la redundancia (RDA) [1] es uno de los muchos métodos posibles para extraer y resumir la variabilidad de un conjunto de variables respuesta que puede ser explicada por un conjunto de variables predictoras. La idea principal es usar regresiones lineales multivariantes para explicar las respuestas en función de las explicativas y luego usar el Análisis de Componentes Principales (ACP) o una representación biplot para visualizar los resultados.

Cuando las variables respuesta son categóricas (binarias, nominales u ordinales), las técnicas lineales clásicas no son adecuadas. Algunas alternativas como RDA basado en distancias han sido propuestos dentro de la literatura. En este trabajo, proponemos una versión del RDA basada en modelos lineales generalizados con respuestas logísticas. Los métodos de visualización naturales para nuestras técnicas son los Biplots Logísticos, recientemente propuestos [2,3].

Los procedimientos serán ilustrados a través de una aplicación a datos reales utilizando el software estadístico R [4] con las funciones desarrolladas dentro del paquete MultBiplotR [5].

Keywords: Logistic Biplot, Redundancy Analysis, Binary Data

References

[1] Van derWollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42(2), 207-2019.

[2] Vicente-Villardon, J. L., Galindo-Villardon, M.P., Blázquez-Zaballos, A. (2006). Logistic biplots. *Multiple correspondence analysis and related methods. Chapman and Hall/CRC, London.* 503-521.

[3] Demey, J. R., Vicente-Villardon, J. L., Galindo-Villardon, M. P., Zambrano, A. Y. (2008). Identifyingmolecular markers associated with classification of genotypes by External Logistic Biplots. *Bioinformatics.*, 24(24), 2832-2838.

[4] R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

[5] Jose Luis Vicente-Villardon (2021). MultBiplotR: Multivariate Analysis Using Biplots in R. R package version 1.3.30. `https://CRAN.R-project.org/package=MultBiplotR`

Sesión 7: Genética e inferencia causal
16 de septiembre, 10:00 a 11:15

Chair: Sofía Aguilar y Blanca Rodríguez

Epigenome-wide study of the exposure to green spaces and blood DNA methylation

Sep 16th
10:00

Sofía Aguilar-Lacasaña[1,2,3], Irene Fontes Marques[4], Serena Fossati[1,2,3], Payam Davdan[1,2,3], Juan R. Gónzalez[1,2,3], Mark J. Nieuwenhuijsen[1,2,3], Mariona Bustamante[1,2,3], Janine Felix[4], Martine Vrijheid[1,2,3]

[1]ISGlobal, Barcelona Spain; [2]Universitat Pompeu Fabra, Barcelona, Spain; [3]CIBER Epidemiología y Salud Pública, Spain; [4]Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands

Previous evidence has shown an association between green spaces exposure and health outcomes. Several mediating pathways have been proposed (physical activity, stress, air pollution, noise, etc.) that might be finally affecting the epigenome. Here we aim to investigate the epigenetic mechanisms that might mediate this association. To this end, we will analyse the association between exposure to green spaces during pregnancy and genome-wide DNA methylation levels in cord blood.

The study has been conducted in 1528 children of the INMA and Generation R European birth cohorts from ATHLETE project. DNA methylation was assessed with the 450K array. To assess exposure to green spaces, we characterized residential surrounding greenness as the average of satellite-based NDVI in buffers of 100m and 300m around residential address. We conducted epigenome-wide association study (EWAS) to identify deferentially methylated CpGs. Associations were estimated using robust linear regression models adjusted by potential confounders. Results from the cohorts were combined through fixed-effects inverse variance weighted meta-analysis.

After multiple-testing correction (FDR), we found significant associations between NDVI within a buffer of 100m and cord blood DNA methylation at two CpG sites (cg13298963, cg01232726) located in ACCN1 and TMPRSS5 genes. No significant associations were observed for NDVI within a buffer of 300m.

Results presented here will be meta-analyzed with results from other 6 cohorts from the LifeCycle and ATHLETE projects. Some of the cohorts will be analyzed through DataShield[1]. The biological interpretation of the results will be done through in silico function enrichment analyses. Also, the effect of life-long exposure to green spaces on child blood DNA methylation will be investigated.

Keywords: Green spaces, DNA methylation, Cord blood

References

[1] A. Gaye et al., "DataSHIELD: Taking the analysis to the data, not the data to the analysis," *Int. J. Epidemiol.*, vol. 43, no. 6, pp. 1929–1944, 2014, doi: 10.1093/ije/dyu188.

Sep 16th
10:15

Renyi divergence measures for the evaluation of surrogate endpoints based on causal inference

Gokce Deliorman[1], Ariel Alonso[2] , Maria del Carmen Pardo[1]

[1]Department of Statistics and O.R., Complutense University of Madrid, Spain; [2]I-BioStat, KU Leuven, Leuven Belgium

In clinical trials, the use of surrogate endpoints can reduce follow-up trial time, the number of required patients and/or the cost. There are many methods for evaluating the surrogate endpoint, and one of them is the causal inference paradigm. Alonso et al. [1] suggested that evaluating surrogacy using individual causal association (ICA) based on mutual information, is defined as the association between the individual causal effects. In addition, when both endpoints are continuous, and normally distributed, the ICA is equivalent to the Pearson correlation coefficient. On one hand, the mutual information between two random variables is equal to the Kullback-Leibler (KL) divergence between the joint probability distribution of these two random variables, and the product of their marginal distributions. On the other hand, KL divergence is a special case of the family of Rényi divergences [2]. Based on this family of divergences, we extended ICA. The performance of the new family is analyzed via a simulation study under normal as well as non-normal scenarios for several sample sizes and ICA levels. Finally, we illustrate the new family of measures to assess surrogacy with a real data set.

Keywords: individual causal association, surrogate endpoints, divergence measures

References

[**1**] Alonso, A., Van der Elst, W., Molenberghs, G., Buyse, M., & Burzykowski, T. (2015). On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics*, 71(1), 15-24.

[**2**] Morales, D., Pardo, L., Pardo, M. C., & Vajda, I. (2004). Rényi statistics for testing composite hypotheses in general exponential models. *Statistics*, 38(2), 133-147.

Exploring quantitative brain features associated with high genetic predisposition to Alzheimer's disease using Compositional Data Analysis

Sep 16th
10:30

Patricia Genius[1,2], Juan D. Gispert, Grégory Operto, Manel Esteller, Arcadi Navarro, Roderic Guigó, Malu Calle, Natalia Vilor-Tejedor

[1]Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation, Barcelona, Spain; [2]Center for Genomic Regulation, Barcelona, Spain.

Imaging genetics studies aim to analyze how genetic information influences brain structure and function by combining neuroimaging-based brain features and genetic data. Most studies focus on standard univariate methods, in which phenotypes are individually analyzed to identify genetic variants associated with them. In the context of high-dimensional data, this strategy translates into reduced statistical power. We propose the application of the selection of balances (Selbal) algorithm [1] a method with higher statistical power based on compositional data analysis (CODA). The main goal is to explore the association between the genetic predisposition to Alzheimer's disease (AD) and the joint modulation of hippocampal brain subregions (target regions for AD).

The sample of the study was defined by 1,071 cognitively unimpaired middle-age participants from the ALzheimer's and FAmilies (ALFA) study with available information on genetics and neuroimaging data. We used the selection of balances (Selbal)-CODA algorithm to assess the joint change of the selected brain components (hippocampal subregions volumes) in relation to the genetic predisposition to AD (Polygenic Risk score of AD). We also defined sex-stratified models to discern sex-specific effects.

Through the application of Selbal, we found that individuals at higher genetic predisposition to AD showed a significant joint volumetric modulation of specific subregions that have not been reported in previous literature, nor in univariate studies. Moreover, we observed sex-differences in the joint modulation of these structures.

This work provides an innovative modelling perspective for IG studies, and emphasizes the need to explore the brain as a composition, providing an approximation that may be closer to the volumetric changes that individuals at higher genetic risk of AD can display.

Keywords: Compositional Data Analysis; Imaging Genetics; Selbal algorithm

References

[**1**] Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., & Calle, M. L. (2018). Balances: a New Perspective for Microbiome Analysis. *MSystems*, 3(4), 1–12. https://doi.org/10.1128/msystems.00053-18

Sep 16th
10:45

Penalized Logistic Regression for Health Status Classification Using Gene Expressions

Carlos J. Peña[1], Juan Carbonell[1]

[1]Unidad de Bioinformática y Bioestadística, Instituto de Investigación Sanitaria de Valencia (INCLIVA)

Logistic regression models are widely used for binary classification by modeling the probability of a certain class. In genomics studies, an important application is the prediction of the current physiological status of cells (such as normal vs diseased) based on gene expression data.

However, gene expression data are quite challenging for Statistics as they belong to what is referred to as high-dimensional data. This sort of data comprises thousands of covariates (i.e. genes) for only a few number of biological samples, resulting in overfitting and multicollinearity issues when traditional statistical methods are applied. Classical statistical approaches are not appropriate in these cases, and other alternatives are needed to avoid overfitting by using less flexible fitting approaches [1].

In this work, we focus mainly on the statistical problems associated with the classification of diseases using gene expression profiles. One of the most important issues is variable selection, also known as feature selection. In this context, there is a large number of genes constantly expressed across the different conditions and, thus, it is important to select those genes that characterize the different health classes. Identifying the most discriminative genes will help improve the classification performance.

Finally, we apply different methods of regularization to a prostate cancer dataset and demonstrate the effectiveness for gene selection in terms of classification accuracy and number of selected genes. The final aim is to choose a subset of useful genes to diagnose the disease among the whole set of genes.

Keywords: penalized regression, gene selection, high-dimensional data

References

[1] Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction. *Springer Series in Statistics.*

A comparison of Mendelian Randomization methods for assessing causal effects on complex traits

Sep 16th
11:00

Blanca Rodríguez-Fernández[1], Juan D. Gispert, Roderic Guigo, Arcadi Navarro, Natalia Vilor-Tejedor, Marta Crous-Bou

[1]Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation.

Background: Observational studies suffer from confounding and reverse causation, which impede to discern whether the association between the exposure and outcome is causal or not. Mendelian Randomization (MR) approaches use genetic variants known to affect risk factors or exposures of interest as surrogates for exposure assessment to overcome problems in traditional epidemiology. Thus, genetic variants serve as instrumental variables to obtain unbiased estimators of the exposure-outcome causal relationship.

Objective: The objective of this study was to compare several robust MR methods to discern causal effects in epidemiological studies. As a proof of concept, we tested the potential causal role of telomere length (TL) (a well-known biomarker of aging) on life expectancy.

Methods: In this study, we used inverse-variance (IVW), MR-Egger regression, weighted median and maximum likelihood methods with summarized data to estimate the causal effect of genetically predicted longer TL on life expectancy. MR-Egger regression intercept-test, leave-one-SNP-out and Cochran Q statistic were used as ad hoc sensitivity analysis for evaluating the robustness of significant results. MR-Pleiotropy RESidual Sum and Outlier (MR-PRESSO) was used to identify horizontal pleiotropic outliers.

Results: Genetically predicted longer TL was significantly associated with increased life expectancy according to IVW method (βIVW = 0.011, SE = 0.004, p-value = 0.010) with no evidence of heterogeneity (Cochran Q p-value $>$ 0.05) nor directional pleiotropy (intercept test p-value $>$ 0.05). Pleiotropy-robust methods (i.e., weighted median and weighted mode) also produced similar patterns of effects, further supporting the robustness of the analyses. There was no evidence of heterogeneity due to SNP outliers according to MR-PRESSO global test. MR-Egger regression method did not support this association (MR-Egger p-value = 0.113).

Conclusions: We recommend utilizing robust methods with different assumptions to test the consistency and robustness of MR results in genetic epidemiology research.

Keywords: Causal inference, epidemiology, genetic association studies

List of participants

In alphabetical order:

№	Name	Institution	e-mail
1.	Sofía Aguilar Lacasaña	ISGlobal	sofia.aguilar@isglobal.org
2.	Laura Aixalà Perelló	Universidad de Valencia (UV)	laixalap@gmail.com
3.	María Anciones Polo	Universidad de Salamanca	mariaanciones@usal.es
4.	Goizalde Badiola Zabala	Universidad del País Vasco (UPV/EHU)	goizalde.badiola@ehu.eus
5.	Irantzu Barrio Beraza	Universidad del País Vasco (UPV/EHU)	irantzu.barrio@ehu.eus
6.	Miguel Ángel Beltrán Sánchez	FISABIO	mianbel@alumni.uv.es
7.	Enrique Benéitez Andrés	Universidad de Salamanca	ebeneitez@usal.es
8.	Raúl Beneyto Menargues	FISABIO	raulbeme@alumni.uv.es
9.	María José Caballero	GVA	marijosecn.21@gmail.com
10.	Gabriel Calvo Bayarri	Universidad de Valencia (UV)	g.calvobayarri@gmail.com
11.	Juan Antonio Carbonell Asíns	INCLIVA	jacarbonell@incliva.es
12.	Jorge Castillo-Mateo	Universidad de Zaragoza	jorgecm@unizar.es
13.	Martina Cendoya Martínez	IVIA	cendoya@alumni.uv.es

№	Name	Institution	e-mail
14.	Irene Creus Martí	Universidad de Valencia (UV)	irenecreus@gmail.com
15.	David Conesa Guillén	Universidad de Valencia (UV)	David.V.Conesa@uv.es
16.	Javier Antonio de la Hoz Maestre	Universidad de Salamanca	jdelahozmaestre@gmail.com
17.	Gokce Deliorman	Universidad Complutense de Madrid	gdeliorm@ucm.es
18.	Pablo Escobar Hernández	Universidad de Valencia (UV)	pablo.escobar@uv.es
19.	Anabel Forte Deltell	Universidad de Valencia (UV)	anabel.forte@uv.es
20.	Alba Fuster Alonso	Universidad de Valencia (UV)	alba.fuster1398@gmail.com
21.	Anaïs Galvañ Domenech	IVIA	galvany_anadom@externos.gva.es
22.	Leire Garmendia Bergés	Universidad Politécnica de Cataluña	leire.garmendia@upc.edu
23.	Patricia Genius Serra	Barcelonaβeta Brain Research Center	pgenius@barcelonabeta.org
24.	Armand González Escalante	Barcelonaβeta Brain Research Center	agonzalez@barcelonabeta.org
25.	Armando González Sánchez	Universidad de Salamanca	armando_gonzalez@usal.es
26.	Harkaitz Goyena Baroja	Universidad Pública de Navarra (UPNA)	harkaitz.goyena@unavarra.es
27.	María Grau Pérez	Universidad Autónoma de Madrid	maria.grau.perez@gmail.com
28.	Gresky Gutiérrez Sánchez	Universidad de Salamanca	greskygutierrez@usal.es
29.	Pavel Hernández Amaro	Universidad Carlos III de Madrid (UC3M)	pahernan@est-econ.uc3m.es
30.	Harold Antonio Hernández Roig	Universidad Carlos III de Madrid (UC3M)	hahernan@est-econ.uc3m.es
31.	Amaia Iparragirre Letamendi	Universidad del País Vasco (UPV/EHU)	amaia.iparragirre@ehu.eus

№	Name	Institution	e-mail
32.	Elena Lázaro Hervás	IVIA	lazaro_ele@gva.es
33.	Fran Llopis-Cardona	FISABIO	llopis_fracar@gva.es
34.	Javier Martín-Pozuelo Lozano	Universidad de Valencia (UV)	martinpo@alumni.uv.es
35.	Joaquín Martínez Minaya	Universitat Politècnica de València (UPV)	jomarminaya@gmail.com
36.	Dorota Mlynarczyk	Universidad Autónoma de Barcelona (UAB)	dori.mlynarczyk@gmail.com
37.	Sergio Moreno Fornés	CEEISCAT	smorenof@iconcologia.net
38.	Alexander Olza Rodriguez	BCAM	aolza@bcamath.org
39.	Jon Aritz Panera Carracedo	UPC-ICO	jonaritzp@gmail.com
40.	Carlos Javier Peña de los Santos	INCLIVA	cpena@incliva.es
41.	Alfonso Ignacio Picó Peris	Universidad de Valencia (UV)	alfonso.i.pico@gmail.com
42.	Marcos Quijal Zamorano	ISGlobal	marcos.quijal@isglobal.org
43.	Garazi Retegui Goñi	Universidad Pública de Navarra (UPNA)	garazi.retegui@unavarra.es
44.	Blanca Rodríguez Fernández	Barcelonaβeta Brain Research Center	brodriguez@barcelonabeta.org
45.	Zulema Rodriguez Hernandez	Centro Nacional de Epidemología (ISCIII)	zulema.rodriguez@isciii.es
46.	Lander Rodríguez Idiazabal	BCAM	lrodriguez@bcamath.org
47.	Cristóbal Manuel Rodríguez Leal	Universidad Complutense de Madrid	cristo07@ucm.es
48.	M Carmen Romero Ferreiro	Instituto de Investigación Biomédica Hospital Universitario 12 de Octubre	marome10@ucm.es

№	Name	Institution	e-mail
49.	Andrea Toloba López-Egea	Universidad Politécnica de Cataluña	andrea.toloba@estudiantat.upc.edu
50.	Antonio Vicent	IVIA	vicent_antciv@gva.es
51.	Laura Vicente González	Universidad de Salamanca	laura20vg@usal.es
52.	Natalia Vilor-Tejedor	CRG & BBRC	natalia.vilortejedor@crg.eu
53.	Lore Zumeta Olaskoaga	BCAM	lzumeta@bcamath.org