CEB
EIB
2023

LIBRO DE RESÚMENES

XIX Conferencia Española y VIII Encuentro Iberoamericano de Biometría

Vigo, del 27 al 30 de junio de 2023

XIX Conferencia Española
y
VIII Encuentro Iberoamericano de Biometría

CEB-EIB 2023

Libro de Resúmenes

Vigo, 27–30 de junio de 2023

XIX Conferencia Española
y VIII Encuentro Iberoamericano de Biometría
(CEB-EIB 2023)

Libro de Resúmenes

Elaborado por:
Juan Carlos Pardo Fernández
María Xosé Rodríguez Álvarez

ISBN: 978-84-8158-974-0

Comité científico

- Presidente: Javier Roca Pardiñas (Universidade de Vigo)

- Danilo Alvares (University of Cambridge, UK)
- Irantzu Barrio (Universidad del País Vasco–Euskal Herriko Unibertsitatea, UPV/EHU)
- M. Luz Calle Rosingana (Universitat de Vic)
- Paulo Canas Rodrigues (Universidade Federal da Bahia, Brasil)
- Mª Eugenia Castellanos (Universidad Rey Juan Carlos)
- Ricardo Cao Abad (Universidade da Coruña)
- Ana Clavería Fontán (Xerencia de atención primaria de Vigo)
- Rosa Crujeiras Casais (Universidade de Santiago de Compostela)
- Maria Durbán (Universidad Carlos III de Madrid)
- Anabel Forte (Universitat de València)
- Liliana Lopez Kleine (Universidad Nacional de Colombia, Colombia)
- Vicente Núñez Antón (Universidad del País Vasco–Euskal Herriko Unibertsitatea, UPV/EHU)
- María Xosé Rodríguez Álvarez (Universidade de Vigo)
- Omar Honorio Ruíz Bartola (Escuela Superior Politécnica del Litoral, Ecuador)
- Inês Sousa (Universidade do Minho, Portugal)
- Silvia Sühring (Universidad Nacional de Salta, Argentina)

Comité organizador

- Presidenta: María Xosé Rodríguez Álvarez (Universidade de Vigo)

- Renata Alcarde Sermarini (Universidade de São Paulo, Brasil)
- Danilo Alvares (University of Cambridge, Reino Unido)
- Sergio Bramardi (Universidad Nacional del Comahue, Argentina)
- Tomás Cotos Yáñez (Universidade de Vigo)
- Jacobo de Uña Álvarez (Universidade de Vigo)
- Diego García Sánchez (Universidade de Vigo)
- Luis Fernando Grajales Hernández (Universidad Nacional de Colombia, Colombia)
- María del Carmen Iglesias Pérez (Universidade de Vigo)
- Juan Carlos Pardo Fernández (Universidade de Vigo)
- Ana Pérez González (Universidade de Vigo)
- Javier Roca Pardiñas (Universidade de Vigo)
- Marta Sestelo (Universidade de Vigo)
- Jessica Martha Vera Bermúdez (Escuela Superior Politécnica del Litoral, Ecuador)

Prefacio

Respondiendo al encargo de la Sociedad Española de Bioestadística (SEB), la XIX Conferencia Española de Biometría (CEB) tiene lugar este año en Vigo. En esta ocasión, la conferencia se celebra conjuntamente con el VIII Encuentro Iberoamericano de Biometría (EIB), y son seis las regiones de la International Biometric Society (IBS) involucradas en la organización: regiones Argentina, Brasileña, Española, Centroamericana y Caribeña, Chilena y Ecuatoriana. La CEB-EIB 2023 ha sido organizada por el Departamento de Estadística e Investigación Operativa y el grupo de investigación SiDOR (Statistical Inference, Decision and Operations Research) de la Universidade de Vigo, y ha contado con la inestimable ayuda de compañeras y compañeros de todas las regiones involucradas.

Organizar una conferencia siempre requiere un esfuerzo grande y son muchas las incertidumbres que hay a lo largo de todo el proceso. Sin embargo, solo podemos estar abrumadas/os y orgullosas/os –aunque conscientes de la responsabilidad que ello implica– por el recibimiento que la CEB-EIB 2023 ha tenido. El programa científico incluye un total de 153 contribuciones, de las cuales 100 se presentan en formato oral y el resto en formato póster. Además, un total de 187 personas asisten a la conferencia. Especial mención merecen las compañeras y compañeros iberoamericanos que nos acompañan en esta ocasión (alrededor del 15 % de los asistentes). Somos conscientes del esfuerzo que supone un desplazamiento intercontinental, y por ello nuestro más sincero agradecimiento.

Un agradecimiento también a las y los miembros del Comité Científico por la labor realizada eligiendo las conferencias plenarias, organizando las sesiones invitadas y seleccionando todos los trabajos que contribuyen al programa y cuyos resúmenes se presentan en este libro. Las conferencias plenarias, impartidas por María Gabriela Cendoya, Andreas Mayr y Anabel Forte Deltell, son del máximo interés, como así lo es la actividad formativa previa a la conferencia, a cargo de María Amalia Jácome y Ricardo Cao. Una mención especial merece la Sesión Especial Iberoamericana, que cuenta con representantes de las regiones organizadoras de la CEB-EIB 2023.

Esta edición viene marcada por una sesión homenaje a Carmen Cadarso Suárez, fallecida repentinamente el 3 de junio del 2022. Carmen fue una bioestadística gallega, catedrática del Departamento de Estadística, Análisis Matemático y Optimización de la Universidade de Santiago de Compostela, coordinadora e impulsora de la Red Nacional de Bioestadística BIOSTATNET y directora del grupo de investigación de Bioestadística y Ciencia de Datos Biomédicos (GRID-BDS), entre otros muchos logros. El legado e influencia de Carmen en el campo de la bioestadística siempre será recordado. Nuestro agradecimiento a Wenceslao González Manteiga, Thomas Kneib y Jenifer Espasandín Domínguez por su participación en esta sesión homenaje.

Como en ediciones anteriores, las y los jóvenes tienen un papel relevante y fundamental. Del total de 153 contribuciones, 68 son de jóvenes investigadoras/es. Hay, además, una Sesión Especial de Jóvenes Investigadoras/es. Todas y todos los jóvenes de las distintas regiones estaban invitadas/os a participar en ella. Nos consta que el nivel de los trabajos enviados fue muy alto y que, por ello, la decisión de quien participaría en la sesión, difícil. Queremos dar la enhorabuena a todas y todos los jóvenes. ¡Es un orgullo contar con una cantera con tanto talento!

Un agradecimiento especial a todos los patrocinadores de la conferencia: a la Universidade de Vigo, por el soporte económico, técnico y logístico; a la IBS por la ayuda económica que permite que cinco jóvenes de Iberoamérica asistan a la conferencia; al Concello de Vigo, que generosamente financia las actividades sociales; a la Deputación de Pontevedra, responsable de la impresión del programa de mano, la cartelería y las acreditaciones; al Departamento de Estadística e Investigación Operativa y al grupo SiDOR, por su apoyo tanto económico como humano.

Finalmente, muchas gracias a todas y todos los asistentes a la CEB-EIB 2023. Sin vosotras/os nada de esto sería posible. Esperamos que disfrutéis de vuestra estancia en Vigo y del programa científico y social de la conferencia.

Benvidas e benvidos á CEB-EIB 2023! Benvidas e benvidos a Vigo!

María Xosé (Coté) Rodríguez Álvarez
Presidenta del Comité Organizador

Vigo, junio de 2023

Este documento contiene los resúmenes de los trabajos presentados en la XIX Conferencia Española y VIII Encuentro Iberoamericano de Biometría (CEB-EIB 2023). El orden de las ponencias de las sesiones especiales es el que aparece en el programa del congreso. Las contribuciones orales y los pósteres están ordenados por orden alfabético del/de la primer/a autor/a. El/la ponente en cada caso aparece indicado en color magenta.

Índice

Conferencias Plenarias

Comparison of different approaches to estimate linkage disequilibrium extent in crop breeding populations

María Gabriela Cendoya (Universidad Nacional de Mar del Plata, Argentina)

The increasing global concern over crop improvement and agricultural production in the face of continuous population growth and climate change has led to the widespread use of genomic selection as a tool in crop breeding. However, traditional, and non-traditional selection methods can result in a significant loss of genetic diversity in modern varieties, making the maintenance of genetic diversity just as important as genetic gain. One way to achieve this is by understanding and measuring linkage disequilibrium (LD), which refers to the non-random association of alleles at different loci. LD is a valuable tool in population genetics and evolutionary biology, used for mapping quantitative trait loci, estimating effective population size and past founder events, and detecting genomic regions under selection. However, measuring the pattern and extent of LD is influenced by several factors, including mating type, genetic drift, gene flow, selection, mutation, population substructure and relatedness, and the statistical tools used. This talk will compare different approaches commonly used to model LD decay and their impact on estimating the extent of LD, focusing on inbred sunflower lines from a mature breeding program.

Statistical boosting for biomedical research: strengths and limitations

Andreas Mayr (Universität Bonn, Alemania)

Biostatisticians nowadays can choose from a huge toolbox of advanced methods and algorithms for prediction purposes. Some of these tools are based on concepts from machine learning; other methods rely on more classical statistical modelling approaches. In clinical settings, doctors are sometimes reluctant to consider risk scores that are constructed by black-box algorithms without clinically meaningful interpretation. Furthermore, even a both accurate and interpretable model will not often be used in practice, when it is based on variables that are difficult to obtain in clinical routine or when its calculation is too complex. In this talk, I will give a non-technical introduction to statistical boosting algorithms which can be interpreted as the methodological intersection between machine learning and statistical modelling. Boosting is able to perform variable selection while estimating statistical models from potentially high-dimensional data. It is mainly suitable for exploratory data analysis or prediction purposes. I will give an overview on some current methodological developments and provide an example for the construction of a clinical risk score. Another example will include the development of polygenic risk scores based on large genetic cohort data.

Past, present and future of Bayesian biostatistics

Anabel Forte (Universitat de València)

The history of Bayesian statistics is deeply linked to biostatistics with one of its first advocates being the statistician Jerome Cornfield, who excelled in the study of cancer and coronary heart

diseases. In general, the data that Biostatistics has been dealing with since its beginnings are complex data that require complex modeling in which Bayesian statistics makes special sense. This situation has not changed over the years and, nowadays, with the so called BigData, it has become even more important to correctly account for the different sources of uncertainty in a problem. In fact, as the future is getting closer we can see the complexity of the data increasing and new approaches arising in order to better understand our world. And in this present and future world, Bayesian Statistics can play an important role not only helping with the inferential process but also in the communication of results, moving away from the conceptual complexity of p-values or confidence intervals.

Curso

Introducción a los modelos de curación

Ricardo Cao, María Amalia Jácome (Grupo de investigación MODES, CITIC, Departamento de Matemáticas, Universidade da Coruña, España)

Introducción a los modelos de curación (*cure models*) desde una visión conceptual y fundamentalmente aplicada. El curso empieza con una pequeña introducción al concepto de cura en Análisis de Supervivencia, y a los principales modelos en presencia de cura. El tema central del curso, los *Mixture Cure Models*, se introducen tanto con un enfoque paramétrico y semiparamétrico como no paramétrico. El curso finaliza con una discusión sobre el abordaje de algunas cuestiones que pueden surgir cuando se aplican modelos de curación, como pueden ser la identificabilidad del modelo y la presencia o no de sujetos curados.

Programa:

1. Modelización de tiempos de vida en presencia de cura
2. Modelos de curación de tipo mixtura
    2.1. Algoritmo EM para modelos de curación de tipo mixtura
    2.2. Modelos de curación de tipo mixtura paramétricos
    2.3. Modelos de curación de tipo mixtura semiparamétricos
    2.4. Modelos de curación de tipo mixtura no paramétricos
3. Estimación de la probabilidad de cura
4. Estimación de la función de latencia
5. Predicción del estado de cura
6. Selección de variables en modelos de curación de tipo mixtura
7. Contrastes de bondad de ajuste
8. Estimación cuando el estado de cura se observa parcialmente
9. Seguimiento suficiente, identificabilidad y estudio de la presencia de individuos curados

Sesión Especial:
Homenaje a Carmen Cadarso

Carmen Cadarso como alumna de doctorado. Aspectos notables de su trabajo de tesis doctoral en la Bioestadística

Wenceslao González-Manteiga (Universidade de Santiago de Compostela, España)

En esta presentación se revisará mi experiencia académica como director de tesis doctoral de la profesora Carmen Cadarso. Sus inicios como alumna de doctorado, muy interesada en el ámbito de la Bioestadística, la proyección de toda su investigación desarrollada en su doctorado en el Análisis de Supervivencia, con los resultados que se derivaron en su tesis doctoral para el estimador de Kaplan-Meier con covariables y sus aplicaciones. Se complementará la presentación con las implicaciones que tuvieron sus estudios en mi vida académica y con las colaboraciones desarrolladas posteriormente.

Una vida para la Bioestadística

Tomas Kneib (Georg-August-Universität Göttingen, Alemania)

Carmen's passion for Biostatistics and international collaboration is well known to all her friends and colleagues. In this presentation, I will reflect on Carmen's relation to Germany in general as well as our shared personal and scientific history. This will include some scientific results but even more memories about all the good times that I could spend with Carmen and her group.

Carmen Cadarso: mi maestra y mentora

Jenifer Espasandín Domínguez (Grupo de Bioestadística y Ciencia de Datos Biomédicos, Universidade de Santiago de Compostela, España)

Investigadora científica, docente, emprendedora y con gran capacidad de liderazgo. Su amplia trayectoria científica de más de 30 años avala su excelencia. A título personal Carmen era muy vital, luchadora, generosa y una apasionada de la cultura y de la música.

Hasta su fallecimiento dirigió el Grupo de Bioestadística y Ciencia de Datos Biomédicos de la Universidade de Santiago de Compostela, en el que yo he tenido la oportunidad de trabajar y realizar mi tesis doctoral. En esta presentación, resumiré mi experiencia como su alumna y algunos de los muchos recuerdos vividos con ella y nuestro grupo de investigación.

Sesión Especial: Iberoamericana

Retos en el análisis de datos de RNAseq de célula única

Liliana López Kleine

llopezk@unal.edu.co, Departamento de estadística. Universidad Nacional de Colombia - sede Bogotá.
Grupo de investigación en Bioinformática y Biología de Sistemas.
Grupo de investigación en Métodos en Bioestadística.
Proyecto de investigación Código hermes: 55049.
Project JAGUAR: mapping immune cell diversity across Latin America.

El análisis de datos de RNAseq de célula única es un área de la estadística genómica reciente que ha tomado como base el análisis de datos de RNAseq de tejido completo (bulk-RNA). Sin embargo, dadas las particularidades de estos datos y el hecho de que se está en la etapa de consolidación de los atlas de células humanas que es, por ejemplo alimentado por el proyecto JAGUAR: mapping immune cell diversity across Latin America, los aspectos más estudiados son actualmente las etapas iniciales de preprocesamiento de dichos datos: 1) Filtros de genes, 2) Reducción de dimensiones, 2) Identificación de grupos de células. En esta presentación se abordarán las experiencias que se han tenido en el grupo de investigación relacionadas con estas etapas de análisis, sus dificultades y perspectivas.

SPEED Stat: an Excel package for teaching experimental statistics

André Mundstock Xavier de Carvalho[1], *Fabrícia Queiroz Mendes*[2]

[1]andre.carvalho@ufv.br, Institute of Agricultural Sciences, Federal University of Viçosa, Brazil
[2]fabricia.mendes@ufv.br, Institute of Agricultural Sciences, Federal University of Viçosa, Brazil

A more multidisciplinary graduate formation is a growing demand in today's world. However, a more multidisciplinary training requires, among other things, a pedagogical effort of objective synthesis of each content. In this sense, when teaching experimental statistics to students of agricultural and biological sciences, the learning curve in software such as R, Python or SAS is slow and consumes valuable time that could be invested in the contents of statistical science itself. Thus, the objective of this work was to develop a simple application in Excel for univariate statistical analysis of experiments. Classic ANOVA procedures (fixed effects) were included for the CRD and RBD models in simple, factorial (double or triple, including additional treatments), split-plot (with or without Satterthwaite correction), split-block, simple nested (mixed model) and repeated measures ANOVA (GG correction) for balanced or slightly unbalanced data. With just one click the tool can run Tukey, SNK, Dunnett, Scott-Knott, Holm or Benjamini-Hochberg tests. In factorials, the program allows executing them with EWER control (except for SNK and Scott-Knott). With 6 clicks the user can also perform regression analysis for linear and non-linear models (including exponential and sigmoidal models). Maximum, minimum, asymptotes or other relevant information from the regression models are also calculated automatically. Non-parametric regression with medians are also facilitated in the program. In addition, the application automatically performs tests for normality, homoscedasticity and additivity requirements. In case of violation, the program also automatically scans for outliers or a suitable transformation. The application also allows the performance of ANCOVA with great ease and provides the Papadakis method for correction of spatial autocorrelation according to a sketch informed by the user. Finally, the application allows you to edit the DF and MSerror values in order to allow other analyses (such as joint analysis, blocks with random effect, among others). Among the main limitations of the application, we can highlight the requirement of good hardware configurations and the limitations of scope of procedures. Important simple procedures such as Pearson's correlation matrix and multiple regression analysis could not be included as they would imply a complete change in the minimalist and intuitive structure of the developed spreadsheet program. However, these procedures are relatively simple to be performed with native Excel functions. The application (available for free at speedstatsoftware.wordpress.com in English, Portuguese and Spanish) has been tested with the resolution of hundreds of examples and will soon be adapted for use in LibreOffice as well.

Keywords: statistical software, statistics teaching, classic experimental designs.

Maximizing Efficiency in Platform Trials with Shared Controls: Optimal allocation strategies

Marta Bofill Roig[1,*], *Ekkehard Glimm*[2], *Tobias Mielke*[3], *Martin Posch*[1]

[1]Section for Medical Statistics, Center for Medical Data Science, Medical University of Vienna
[2]Advanced Methodology and Data Science, Novartis Pharma AG, Basel
[3]Statistics and Decision Sciences, Janssen-Cilag GmbH
*marta.bofillroig@meduniwien.ac.at

Platform trials are randomized clinical trials that allow simultaneous comparison of multiple interventions, usually against a common control. Arms to test experimental interventions may enter and leave the platform over time. Therefore, the number of experimental intervention arms in the trial can change over time. Determining optimal allocation rates to allocate patients to the treatment and control arms in platform trials is challenging because the change in treatment arms implies that also the optimal allocation rates will change when treatments enter or leave the platform. In addition, the optimal allocation depends on the analysis strategy used.

In this talk, we describe optimal treatment allocation rates for platform trials with shared controls, assuming that a stratified estimation and testing procedure based on a regression model is used to adjust for time trends. We consider analysis methods using concurrent controls only as well as methods based on also non-concurrent controls. Assuming that the objective function to be minimized is the maximum of the variances of the effect estimators, we show that the optimal solution depends on the entry time of the arms in the trial and, in general, does not correspond to the square root of k allocation rule used in the classical multi-arm trials. We illustrate the optimal allocation and evaluate the power and type 1 error rate compared to trials using one-to-one and square root of k allocations by means of a case study.

Keywords: Clinical trials, Optimisation, Shared controls.

Modelación estadística y aprendizaje automático para predicción genómica en mejoramiento genético vegetal

Mónica Balzarini[1]

[1]mbalzari@gmail.com, Cátedra de Estadística y Biometría. Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba. Unidad de Fitopatología y Modelización Agrícola - Consejo Nacional de Investigaciones Científicas y Técnicas (UFYMA INTA-CONICET), CCT CONICET Córdoba, Argentina.

El mejoramiento genético de caracteres cuantitativos en vegetales es complejo y lento. Los genotipos son evaluados en múltiples años y sitios debido al alto impacto del ambiente. La selección genómica (SG) permite incrementar la tasa de ganancia genética mediante la integración de datos fenotípicos y genómicos en poblaciones de entrenamiento y su posterior uso en la predicción del valor estimado de mejora (GEBV) sólo desde información genómica. La predicción del GEBV se realiza con gran cantidad de datos moleculares; consecuentemente, muchas de las variables predictoras son, a menudo, irrelevantes y redundantes. El objetivo de este trabajo es comparar procedimientos para ajustar modelos de SG con vasta cantidad de marcadores moleculares. Cuatro métodos estadísticos para ajustar modelos predictivos (Bayes A, Bayes B, Bayes C y Regresión Ridge) más Random Forest como algoritmo de aprendizaje automático fueron comparados, con y sin selección previa de marcadores basada en la identificación de marcadores significativos con GWAS (Genome Wide Association Studies). Los modelos se entrenaron con dos poblaciones de genotipos de caña de azúcar (Sacharum spp.) y en un panel diverso de genotipos de maíz (Zea mays L.) con gran cantidad de datos moleculares y evaluaciones fenotípicas multi-ambientales. La eficiencia fue evaluada a través de la correlación entre los GEBV predichos y los BLUP de los efectos genotípicos luego de descontar efectos ambientales desde los datos fenotípicos. Se observó similar eficiencia entre los métodos de ajuste del modelo de SG, en todas las poblaciones. La eficiencia de los modelos estadísticos de SG se incrementó considerablemente con la selección previa de marcadores por GWAS. La precisión de las predicciones augura la selección de genotipos desde información molecular.

Keywords: Marcadores Moleculares, Modelos, Predicción genómica, Eficiencia.

A two-stage approach for Bayesian joint models: Reducing complexity while maintaining accuracy

Danilo Alvares[1], *Valeria Leiva-Yamaguchi*[2]

[1]danilo.alvares@mrc-bsu.cam.ac.uk, MRC Biostatistics Unit, University of Cambridge
[2]vjleiva@mat.uc.cl, Department of Statistics, Pontificia Universidad Católica de Chile

Several joint models for longitudinal and survival data have been proposed in recent years. In particular, many authors have preferred to employ the Bayesian approach to model more complex structures, make dynamic predictions, or use model averaging. However, Markov chain Monte Carlo methods are computationally very demanding and may suffer convergence problems, especially for complex models with random effects, which is the case for most joint models. These issues can be overcome by estimating the parameters of each submodel separately, leading to a natural reduction in the complexity of the joint modelling, but often producing biased estimates. Hence, we propose a novel two-stage approach that uses the estimations from the longitudinal submodel to specify an informative prior distribution for the random effects when estimating them within the survival submodel. In addition, as a bias correction mechanism, we incorporate the longitudinal likelihood function in the second stage, where its fixed effects are set according to the estimation using only the longitudinal submodel. Based on simulation studies and real applications, we empirically compare our proposal with joint specification and standard two-stage approaches considering different types of longitudinal responses (continuous, count and binary). The results show that our estimator is more accurate than its two-stage competitor and as good as jointly estimating all parameters. Moreover, the novel two-stage approach significantly reduces the computational time compared to the joint specification.

Keywords: Bias reduction, Longitudinal data, Time-to-event.

Sesión Especial:
Jóvenes Investigadoras e Investigadores

Joint quantile autoregressive modeling for univariate and spatial time-series data in a Bayesian framework

Jorge Castillo-Mateo[1], *Alan E. Gelfand*[2], *Jesus Asin*[1], *Ana C. Cebrian*[1]

[1]jorgecm@unizar.es, jasin@unizar.es, acebrian@unizar.es, Department of Statistical Methods, University of Zaragoza

[2]alan@stat.duke.edu, Department of Statistical Science, Duke University

Abstract

We propose a fully Bayesian joint quantile autoregression (QAR) modeling for time-series data. We derive a characterization of the noncrossing QAR(1) model using two monotone curves. We offer novel metrics to assess the adequacy of the QAR. Subsequently, we propose a novel spatial joint QAR for spatially referenced time-series data. We illustrate the models with an analysis of persistence in daily maximum temperature data collected in Aragón, Spain.

Keywords: Bayesian methods, daily temperature persistence, Gaussian copula process

1. Introduction

Quantile regression (QR) offers a flexible tool to capture changing explanation across quantile levels between the response and the covariates. The usual approach is the so-called *multiple* QR [1, 2], fitting a separate regression for each quantile of interest, leading to the possibility of crossing of the regression across quantiles. The approach called *joint* QR [5] avoids quantile crossing over a restricted support for the covariates. On the other hand, a seminal version of a joint quantile autoregression (QAR) model was proposed by Koenker and Xiao [4] (KX2006, hereafter). They required all the coefficients of the autoregression to be comonotonic, this is, strictly increasing functions across quantile levels $\tau \in (0, 1)$. Our contribution here is to reconsider the work by KX2006 in the context of Tokdar and Kadane [5] (TK2012, hereafter) to propose novel joint QAR modeling in a Bayesian framework with greater flexibility than KX2006. Going one step further, in the spatial setting, we introduce spatial dependence in the time-series realizations but as well, we add spatially-varying coefficients in order to obtain spatially-varying QAR's, generalizing the spatial QR model by [3] in the sense that they capture spatial dependence through a copula process but obtain a common quantile function that does not vary spatially.

2. Joint QAR model for time-series data

The support of the data. Let $\{y_t^* : t = 1, \ldots, T\}$ be the time-series data. For a noncrossing QAR(1) specification interest focuses on ensuring that the quantile curves do not cross for all values of y_{t-1}^* in a bounded interval. Although the region of interest for noncrossing must be assumed to be bounded,

the variable space itself may still be unbounded. If noncrossing were desired in QAR on an unbounded domain, the result will be parallel lines, yielding the constant autoregression model. We take this interval to be $[0, 1]$ and implement this by making a transformation of the data, $y_t = (y_t^* - m)/(M - m)$, where $m < \min y_t^*$ and $M > \max y_t^*$. For a convenient "automatic" strategy for selecting m and M we use basic results from the theory of order statistics where $y^*_{(1)}$ is the minimum and $y^*_{(T)}$ is the maximum of the data. We propose $m = (Ty^*_{(1)} - y^*_{(T)})/(T - 1)$ and $M = (Ty^*_{(T)} - y^*_{(1)})/(T - 1)$.

The model. A straightforward characterization of the required monotonicity of the QAR(1) lines is:

Theorem 1. *An autoregressive specification, $Q_{Y_t}(\tau \mid y_{t-1}) = \theta_0(\tau) + \theta_1(\tau)y_{t-1}$ with $\theta_1(\tau) \in [-1, 1]$ for $\tau \in [0, 1]$, is monotonically increasing in τ for $y_{t-1} \in [0, 1]$ if and only if $Q_{Y_t}(\tau \mid y_{t-1}) = \eta_2(\tau) + (\eta_1(\tau) - \eta_2(\tau))y_{t-1}$ where $\eta_1, \eta_2 : [0, 1] \to [0, 1]$ are monotonically increasing.*

A model for functions η_1 and η_2 induces a QAR(1) model over all valid QAR(1) specifications of $Q_{Y_t}(\tau \mid y_{t-1})$, provided the boundary conditions $Q_{Y_t}(0 \mid y_{t-1}) = 0$ and $Q_{Y_t}(1 \mid y_{t-1}) = 1$ for all $y_{t-1} \in [0, 1]$ are satisfied, or equivalently, $\eta_j(0) = 0$ and $\eta_j(1) = 1$ $(j = 1, 2)$. A convenient class of η's to work with are cdf's for continuous random variables with support $[0, 1]$. In fact, a rich class would arise as probabilistic mixtures of such cdf's, leading to the general form $\eta(\tau) = \sum_{k=1}^{K} \lambda_k F(\tau \mid \mathbf{\Omega}_k)$, such that $\lambda_k \geq 0$, $\sum_k \lambda_k = 1$ and $F : [0, 1] \to [0, 1]$ is strictly increasing for any parameters $\mathbf{\Omega}_k$. A convenient class of F's are the cdf's of the two parameter Kumaraswamy distribution. This cdf is $F(x \mid a, b) = 1 - (1 - x^a)^b$ where $x \in [0, 1]$ and $a, b > 0$. The Kumaraswamy distributions are a family with behavior similar to the beta distribution but much simpler, especially in the context of simulation since the cdf can be expressed in closed form. Through simulation, we explored that $K = 1$ and $K = 2$ offer great flexibility and a higher K can lead to identification issues. We call these models QAR1K1 and QAR1K2, respectively. We conclude the model specification with the prior distribution of the parameters a's, b's, and λ's. We suggest to model the weights using the additive logistic normal transformation and the parameters of the Kumaraswamy distribution with a weak Gaussian prior in the log scale.

Likelihood evaluation and model fitting. Following the ideas of TK2012, a valid joint specification of $Q_{Y_t}(\tau \mid y_{t-1})$ for all $\tau \in (0, 1)$ uniquely defines the conditional response density for $y_{t-1} \in [0, 1]$,

$$f_{Y_t}(y_t \mid y_{t-1}) = \left. \frac{1}{\frac{d}{d\tau} Q_{Y_t}(\tau \mid y_{t-1})} \right|_{\tau = \tau_{y_{t-1}}(y_t)}, \tag{1}$$

where $\tau_{y_{t-1}}(y_t)$ solves $y_t = y_{t-1}\eta_1(\tau) + (1 - y_{t-1})\eta_2(\tau)$ in τ and is numerically approximated via a one-dimensional rootfinder. Consequently, given y_1, we can write a valid log-likelihood score in terms of $u_t = \tau_{y_{t-1}}(y_t)$, all of the observed data $\mathbf{y} = (y_1, \ldots, y_T)^\top$ and the model parameters $\mathbf{\Omega}$ as

$$\ell(\mathbf{\Omega} \mid \mathbf{y}) = -\sum_{t=2}^{T} \log \left\{ y_{t-1}\dot{\eta}_1(u_t) + (1 - y_{t-1})\dot{\eta}_2(u_t) \right\}. \tag{2}$$

The rootfinder used to evaluate the log-likelihood function (2) is Brent's method. We implement an adaptive block-Metropolis sampler algorithm to obtain Markov chain Monte Carlo (MCMC) samples from the posterior distribution of the parameters and the conditional quantile function.

Model adequacy and comparison. We offer two novel dimensionless metrics which assess the global adequacy and comparative performance of the conditional quantile function arising under the model. They are based on the posterior distribution of $Q_{Y_t}(\tau \mid y_{t-1})$. The first metric $\tilde{p}_v$ uses the probability that an observation is less than each conditional quantile. The second metric $\bar{R}^1$ is a generalization of $R^1(\tau)$, the analog of R^2 for the quantile loss function.

3. Joint spatial QAR model for spatio-temporal data

We focus on the analysis of spatial point-referenced time-series data where $Y_t(\mathbf{s})$ denotes the observation for time $t = 1, \ldots, T$ at location $\mathbf{s} \in \mathcal{D}$, where $\mathcal{D} \subset \mathbb{R}^r$ is the study region. The joint spatial QAR model is given by

$$Y_t(\mathbf{s}) = \theta_0(U_t(\mathbf{s}); \mathbf{s}) + \theta_1(U_t(\mathbf{s}); \mathbf{s}) Y_{t-1}(\mathbf{s}), \tag{3}$$

where the θ functions are quantile and spatially varying, and the vectors $(U_t(\mathbf{s}_1), \ldots, U_t(\mathbf{s}_n))^\top$ are assumed to follow a spatial copula process.

Modeling spatial dependence. *Spatially varying quantiles.* For the spatially-varying coefficients, we consider one cdf for each $\eta(\tau; \mathbf{s})$. In fact, at location $\mathbf{s}$, let assume $\eta_j(\tau; \mathbf{s}) = 1 - (1 - \tau^{a_j(\mathbf{s})})^{b_j(\mathbf{s})}$ with parameters $a_j(\mathbf{s})$ and $b_j(\mathbf{s})$ $(j = 1, 2)$. We introduce four independent GP's for the a's and b's on the log scale. In particular, we model $\log a_j(\mathbf{s}) \sim GP(a_j, \sigma^2_{a_j} \rho(\mathbf{s}, \mathbf{s}'; \phi_{a_j}))$ and $\log b_j(\mathbf{s}) \sim GP(b_j, \sigma^2_{b_j} \rho(\mathbf{s}, \mathbf{s}'; \phi_{b_j}))$ where $\rho(\mathbf{s}, \mathbf{s}'; \phi)$is an exponential correlation functions with decay ϕ.

The spatial copula process. With regard to the copula model for (3), we take the processes $U_t(\mathbf{s})$'s to follow a Gaussian copula for each t, induced by a stationary spatial GP. In the spirit of [3], we define

$$U_t(\mathbf{s}) = \Phi(Z_t(\mathbf{s})), \quad Z_t(\mathbf{s}) = W_t(\mathbf{s}) + \epsilon_t(\mathbf{s}), \quad W_t(\mathbf{s}) \sim GP(0, \gamma\rho(\mathbf{s}, \mathbf{s}'; \phi)), \quad \epsilon_t(\mathbf{s}) \sim \text{IID } N(0, 1-\gamma). \tag{4}$$

The process $W_t(\mathbf{s})$ captures spatial dependence while $\epsilon_t(\mathbf{s})$ is independent pure error. The parameter $\gamma \in [0, 1]$ determines the proportion of spatial and independent variation. With this approach, the Gaussian copula density has correlation matrix $R \equiv \gamma R(\phi) + (1-\gamma)\mathbf{I}_n$ where $R(\phi)$ is the $n \times n$ correlation matrix induced by $\rho(\mathbf{s}, \mathbf{s}'; \phi)$.

Likelihood evaluation and spatial interpolation. We are interested in the likelihood under model (3) and (4). It is convenient to first obtain the joint distribution for all data, $\mathbf{y}$. By Sklar's theorem, the joint conditional density of $\mathbf{y}$ can be partitioned into a marginal part and a copula part. Subsequently, we find the expression of the log-likelihood function for the spatial QAR, and after giving weakly informative priors, inference proceeds in a similar way as in the univariate case. With the proposed model we can interpolate conditional quantiles to any desired location in the study region given any proposed or reference value for the previous day's temperature at that location.

4. Application to Temperature Data

The analyses consider daily maximum temperature (°C) data at $n = 18$ sites around the Comunidad Autónoma de Aragón provided by the Agencia Estatal de Meteorología (AEMET) in northeastern Spain. We use data at a daily scale in 2015, but we focus the analyses on the warm months from May 1 to September 30.

Illustratively, Figure 1 shows the posterior mean of the functions θ_0 and θ_1 in Zaragoza for the

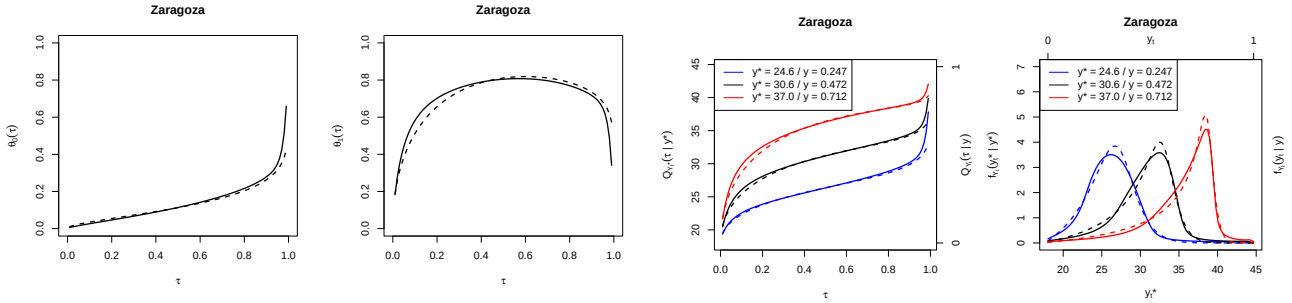


Figure 1: Posterior mean of (1st) $\theta_0(\tau)$, (2nd) $\theta_1(\tau)$, (3rd) quantile function $Q_{Y_t}(\tau \mid y)$ vs. τ; and (4th) density function $f_{Y_t}(x \mid y)$. y is the empirical τ-marginal quantile, $\tau = 0.1$ (blue), 0.5 (black), 0.9 (red).

models QAR1K1 (dashed) and QAR1K2 (solid). The intercepts on the original scale can be recovered as $\theta_0^*(\tau) = m(1 - \eta_1(\tau)) + M\eta_2(\tau)$. Mainly, note that θ_1 is nonmonotonic with smaller values in the extremes, this means that the previous day's temperature is less influential for extreme quantiles. This characteristic in the persistence of temperature was observed in [1, 2]. It cannot be reproduced by KX2006. Although higher K offers more flexibility, both curves offer similar results. Additionally, Figure 1 shows the posterior mean of the conditional quantile functions $Q_{Y_t}(\tau \mid y)$ for three situations where y is the empirical τ-marginal quantile for $\tau = 0.1, 0.5, 0.9$. The figure also shows the posterior mean of the conditional density function in (1) under the same conditions. The shape of the distribution changes according to the value on which we condition.

The spatial QAR model is fitted to the $n = 18$ series jointly. The posterior mean of γ, the proportion of spatial dependence in (4), is 0.95 with a 95% credible interval of $(0.93, 0.97)$. This result indicates very strong spatial dependence in the quantile levels. Results about spatial GP's for the parameters of the Kumaraswamy cdf (not shown) suggest that the GP of $a_2(\mathbf{s})$ might be not necessary but the spatial variability of $a_1(\mathbf{s})$ is higher and it could be related to distance to coast. We notice that $b_1(\mathbf{s})$ and $b_2(\mathbf{s})$ show approximately negative spatial correlation against each other because $b_1(\mathbf{s})$ has the highest values where $b_2(\mathbf{s})$ has the lowest.

5. Extensions and Future Work

The complete work also includes an approach for the QAR(p) case and a novel multivariate QAR for multivariate time-series data using a copula process. A future direction will consider a proper implementation of covariates in the joint QAR setting. Another interesting direction is to build a bivariate spatial QAR model for daily maximum and minimum temperature.

6. Acknowledgments

This work was partially supported by the Ministerio de Ciencia e Innovación under Grant PID2020-116873GB-I00 and TED2021-130702B-I00; Gobierno de Aragón under Research Group E46_20R: Modelos Estocásticos; and J. C.-M. was supported by Gobierno de Aragón under Doctoral Scholarship ORDEN CUS/581/2020. This work was done while J. C.-M. was a Visiting Scholar at Duke University.

7. Bibliography

[1] Castillo-Mateo J., Asín J., Cebrián A. C., Gelfand A. E. and Abaurrea J. (*in press*). Spatial quantile autoregression for season within year daily maximum temperature data. *Annals of Applied Statistics*. https://doi.org/10.1214/22-AOAS1719

[2] Castillo-Mateo J., Gelfand A. E., Asín J. and Cebrián A. C. (2022). Spatio-temporal quantile autoregression for detecting changes in daily temperature in northeastern Spain. In S. Cabras, I. Cascos, M. E. Castellanos, M. Durbán (Eds.), *Book of Abstracts XVIII Congreso de Biometría CEB-MADRID* (pp. 122–126). Universidad Carlos III de Madrid.

[3] Chen X. and Tokdar S. T. (2021). Joint quantile regression for spatial data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4), 826–852.

[4] Koenker R. and Xiao Z. (2006). Quantile autoregression. *Journal of the American Statistical Association*, 101(475), 980–990.

[5] Tokdar S. T. and Kadane J. B. (2012). Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Analysis*, 7(1), 51–72.

`knobi`: an R package for estimating effects of environmental variability on the fish stocks production

Anxo Paz[1], *Marta Cousido-Rocha*[2], *Maria Grazia Pennino*[3], *Santiago Cerviño*[4]

[1]anxo.paz@ieo.csic.es, [2]marta.cousido@ieo.csic.es, [4]santiago.cervino@ieo.csic.es,
Instituto Español de Oceanografía (IEO - CSIC), Centro Oceanográfico de Vigo, Subida a Radio Faro, 50-52, 36390, Vigo

[3]grazia.pennino@ieo.csic.es, Instituto Español de Oceanografía (IEO - CSIC), Sede Central, Calle del Corazón de María, 8 28002, Madrid

Abstract

In order to understand the population dynamics of marine resources, science had to come up with large collection of statistical and mathematical models, known as stock assessment models, however, most of them ignore the possible environmental effects and biological interactions on the stocks dynamics. Known biomass production models (KBPMs), implemented in the `knobi` R package, provide a useful and simple tool to model overall productivity considering that it responds in a non-linear way to multiple drivers related to climatic, anthropogenic and ecological factors.

Keywords: Stock assessment model, R package, non-linear models.

1. Introduction

Living marine resources inhabit complex ecosystems, influenced by multiple drivers that operate and interact at multiple scales that can result in non-linear or abrupt responses to disturbances. Mathematical and statistical techniques (called assessment methods) can be applied to understand the impact of fishing and environmental effects on fish stocks. Depending on the available data the options goes from simple data-limited methods to more complex age or length-structured methods. However, the majority of models ignore the different environmental effects and biological interactions, hence stock assessment models that can address this issue in a simple way are needed.

The known-biomass production models (KBPMs), introduced by MacCall in 2002 ([1]) are addressed in this study as a useful and simple tool to model overall productivity considering that it responds in a non-linear way to multiple drivers related to climatic, anthropogenic and ecological factors. In particular, KBPMs can identify possible changes in fish stock production caused by different reason as, for example, the environmental variability.

For this reason, we have developed `knobi` package, which implements KBPMs in a simple and intuitive way allowing to assess the stock productivity of relevant fishery resources.

2. Methods

In fisheries research, surplus production (SP) is defined as the change in stock size that would have taken place if there had been no exploitation. KBPMs are based on the idea that the annual surplus production, in an unfished stock, is equal to the difference in biomass for two consecutive years, and that, for a fished stock, the calculation of surplus production depends on catch.

$$SP_t = \overline{B}_{t+1} - \overline{B}_t + C_t \tag{1}$$

where SP_t is the SP in the year t, $\overline{B}_t$ is the average spawning stock biomass (SSB) in year t, i.e. $\overline{B}_t = (B_t + B_{t+1})/2$, being B_t is the stock SSB at the beginning of the year t, and C_t represent the catch in year t.

KBPMs use as input data a SSB time series produced by other more complex stock assessment model. Then, surplus production is calculated from the known average biomass (of two consecutive years) and the observed catch using equation 1. The KBPMs fit the observed data, SSB and SP, to the production curve in (refec:3) using the quasi-Newton optimization method L-BFGS-B. The formulation in (refec:3) is derived from the model defined by Pella and Tomlinson in 1969 ([2]),

$$SP_t = \frac{r}{p}\overline{B}_t \left(1 - \left(\frac{\overline{B}_t}{K}\right)^p\right) \tag{2}$$

where r is the intrinsic population grown rate, K is the virgin biomass and p is the asymmetry parameter, used so that the production curve is not always symmetrical.

As mentioned above, KBPMs offer the possibility of considering environmental variability. In `knobi` these effects are included as additive and multiplicative effects in the base formulation 2. Then, the additive and multiplicative models are, respectively,

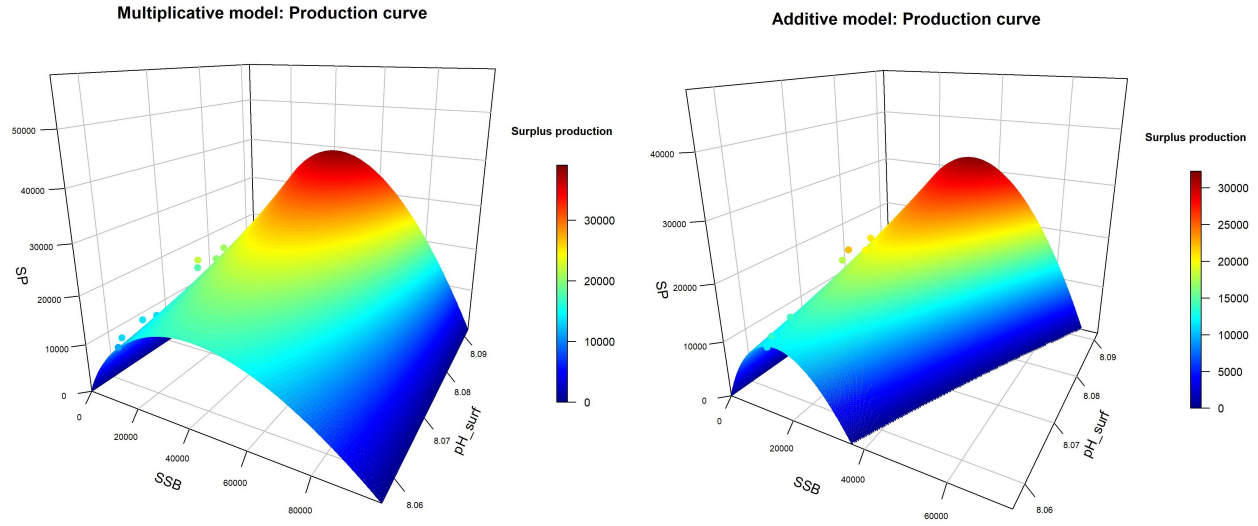
$$SP_t = \frac{r}{p}\overline{B}_t \left(1 - \left(\frac{\overline{B}_t}{K}\right)^p\right) + cX_t\overline{B}_t; \qquad SP_t = \frac{r}{p}\overline{B}_t \left(1 - \left(\frac{\overline{B}_t}{K}\right)^p\right) exp^{cX_t} \tag{3}$$

being c the parameter that represents the effect of the environmental variable X_t, where t represents time (years).

From the equations 2 and 3, the so-called reference points can be obtained, which are informative quantities on the state of exploitation of the different stocks, such as the maximum sustainable yield (MSY) or its associated biomass and fishing mortality.

3. `knobi` package

In this section, `knobi` use and abilities are illustrated through the case study of the European hake (*Merluccius merluccius*) southern stock, which is a demersal species found in the southern Bay of Biscay



(a) Multiplicative environmental KBPM (b) Additive environmental KBPM

Figure 1: `knobi_env` example. Production curves according to the environmental variable values for (a) multiplicative model and (b) additive model in the European hake (southern stock) case study. The environmental variable is *pH_surf* (surface pH).

and Atlantic Iberian waters.

The first step is the model fit through the `knobi_fit` function using the spawning stock biomass (SSB) time series, the catch time series and the corresponding years. Once the KBPM fit is carried out, the `knobi_retro` function tests its robustness to the deletion of data. After that, the environmental effects over the surplus production can be addressed through the `knobi_env` function. More precisely, the function analyse and model the relationships between the surplus production and the environmental variables (Figure 1), implementing a variable selection process based on their correlation with the residuals of a KBPM model that only considers the SSB of the stock as a covariable. Alternatively, the entire set of environmental variables can also be considered in the adjustment instead of carrying out the variable selection process.

Finally, `knobi_proj` function allows us to check the future evolution of the biomass and production of the stock depending on the environmental scenario and the assumed fishing pressure (Figure 2). Then, it allows us to analyze the future status of the stock under different possible settings of fishing pressure or, for example, under the Representative Concentration Pathway (RCP) scenarios defined by the Intergovernmental Panel on Climate Change (IPCC).

4. Conclusions

`knobi` package implements for the first time the known biomass production models, providing a powerful tool for analyzing the stock status from a surplus production point of view. Additionally, the package illustrates KBPMs potential and use, and highlights their advantages. In particular, KBPMs simplicity facilitates the consideration of important drivers that influence the stocks dynamics as the

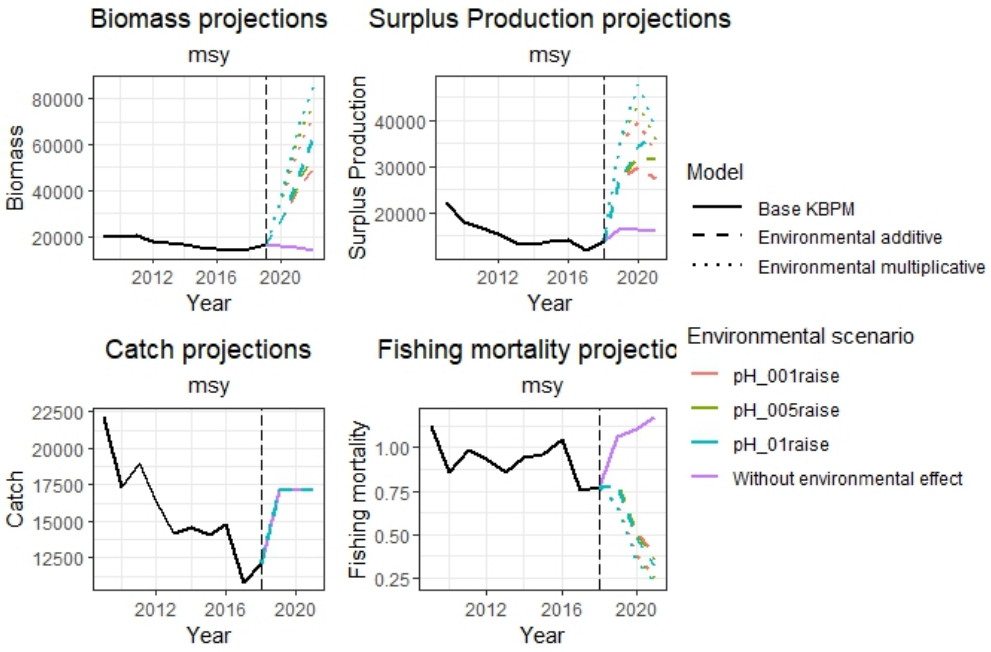


Figure 2: `knobi_proj` example. Each model projections for different future values of the environmental variable surface pH when catches are equal to the Maximum Sustainable Yield (MSY) obtained from the KBPM fit considering only the biomass covariable.

climate change.

For the correct understanding of the package use, `vignettes` in the package help at `https://github.com/MERVEX-group/knobi` are available including illustrative examples.

5. Acknowledgements

Proyecto financiado por la Unión Europea-NextGenerationEU. Componente 3. Inversión 7. Convenio entre el Ministerio de Agricultura, Pesca, Y Alimentación y la Agencia Estatal Consejo Superior de Investigaciones Científicas M.P. -A Través del Instituto Español de Oceanografía- Para impulsar la investigación pesquera como base para la gestión pesquera sostenible. Eje4, FishClim: Conocimiento científico para la adaptación al cambio climático del sector pesquero español

6. Bibliography

[1] MacCall A. (2002). Use of Known-Biomass Production Models to Determine Productivity of West Coast Groundfish Stocks. *North American Journal of Fisheries Management*, 22, 272-279.

[2] Pella J.J., Tomlinson P.K. (1969). A generalized stock-production model. *Bulletin of the Inter-American Tropical Tuna Commission*, 13, 421–58.

Microbiome compositional data analysis for survival studies

Meritxell Pujolassos[1], *Antoni Susín*[2], *M.Luz Calle*[3]

[1]meritxell.pujolassos@uvic.cat, Bioscience Department, Faculty of Sciences, Technology and Engineering, University of Vic – Central University of Catalunya, Vic, Spain

[2] toni.susin@upc.edu, Mathematical Department, UPC-Barcelona Tech, Barcelona, Spain

[3]malu.calle@uvic.cat, Bioscience Department, Faculty of Sciences, Technology and Engineering, University of Vic – Central University of Catalunya, Vic, Spain

Abstract

The compositional nature of microbiome data requires specific compositional data analysis (CoDA) methods. We present a new methodology for the identification of microbial signatures in time-to-event studies. The algorithm implements a CoDA adaptation of elastic-net penalized Cox regression and is integrated in the R package *coda4microbiome* as an extension of the existing functions for cross-sectional and longitudinal studies.

Keywords: microbiome, compositional data analysis, survival data

1. Introduction

Human microbiome is the complete set of microbes found in our bodies, and it plays an important role on human health. Challenging experimental and computational analysis are required to investigate the presence of different microorganisms and understand the complex interactions between them and the environment. High throughput sequencing techniques (16S rRNA and shotgun sequencing), used for identification and quantification of microbial communities, have a limited sequencing capacity which limits the total number of reads that can be revealed from the sample. This total sum restriction implies a great dependence between bacterial species in the analysed sample [1]. Data constraint to a total sum is called compositional data. Compositions are vectors of real positive numbers that contain relative information, which means that each part of the composition on its own is not informative. Information of a composition is extracted from the relation between two or more components [2]. Microbiome data is compositional, therefore, performing its analysis with methods that do not take in account its compositionality may lead to inaccurate results [3], [4].

Compositional Data Analysis (CoDA) was stablished by Aitchison in 1982 [5], who introduced the so-called log-ratio approach, that consists of analysing logarithms of ratios between components instead of each component separately.

coda4microbiome [6] is an algorithm for microbiome analysis based on the log-ratio approach that aims to find a model (microbial signature) with the highest prediction accuracy. *coda4microbiome* has been implemented as an R package and it is developed for cross-sectional and longitudinal studies. In this work, we present a new methodology that extends *coda4microbiome* algorithm to survival data.

The new approach implements elastic-net penalized Cox regression conveniently adapted to CoDA to identify a set of microbial species, and maybe other variables, associated to survival time, i.e., the time until the occurrence of an event of interest, such as, disease onset, response to a treatment, remission, or death.

2. Methods

coda4microbiome algorithm is developed to characterize a microbial signature that best predicts the response variable, and it is structured in three main steps: modelling, variable selection and reparameterization. (1) A regression model with all pairwise log-ratios of microbial species is considered (modelling step), followed by (2) a variable selection step with elastic-net penalization that identifies those log-ratios more associated to the outcome; finally (3) the linear predictor of the log-ratio model is reparametrized to obtain a microbial signature written in terms of the selected bacteria, instead of pairs of bacteria (reparameterization step). Bellow we describe the new *coda4microbiome* algorithm for survival studies.

Assume a survival study with *n* subjects where the time when the event of interest occurs for subject *i* is denoted as t_i. Let $X_i = (X_{i1}, X_{i2}, \dots, X_{iK})$ be the microbial composition for *K* taxa in the *i*-th subject. Microbial abundances (*X*) can be either raw counts or relative abundances. The goal of this method is to identify those microbial taxa whose relative abundances are associated to survival time.

We consider the Cox's proportional hazard regression model (1970) [7] with all possible pairwise log-ratios of taxa as covariates (1). This regression model finds the relationship between pairs of microbes (log-ratios) and the risk of the given event to occur. Using log-ratios in the model, CoDA's principal of scale invariance is ensured.

$$h(t|x) = h_0(t) \cdot \exp\left(\sum_{1\le j<k\le K} \beta_{jk} \cdot \log\left(X_j/X_k\right)\right) \quad (1)$$

Variable selection is carried out by the estimation of the regression coefficients (β_{jk}) subjected to an elastic-net penalization (2) where *L* is the loss function for (1). This step allows the removal of those log-ratios less associated to the outcome, thus only log-ratios with non-zero coefficients are kept. Such penalization can also be written in terms of λ and α, which control the amount of penalization and the mixing between norms, respectively (3).

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\{L(\beta) + \lambda_1\|\beta\|_2^2 + \lambda_2\|\beta\|_1\} \quad (2)$$

$$\lambda_1 = \lambda(1-\alpha);\ \lambda_2 = \lambda\alpha \quad (3)$$

By default, α is set to 0.9 (adjustable by the user), and optimal λ value is selected after a cross-validation process performed by *cv.glmnet()* from glmnet R package [8]. *coda4microbiome* also allows non-compositional variables (i.e., age, sex, clinical variables, etc) to adjust the model.

After modelling and variable selection, the result is a Cox model composed by the logarithms of pairs of bacteria with non-zero coefficient that are associated to the outcome. The linear term of the model, the linear predictor (right side of equation 4), provides an individual prediction (microbial signature score) for the survival time.

$$M_i = \log\left(h_i(t)/h_0(t)\right) = \sum_{1\le j<k\le K} \hat{\beta}_{jk} \cdot \log\left(X_j/X_k\right),\ i\in\{1,\dots,n\} \quad (4)$$

The linearity of logarithms permits the reparameterization of (4) into single bacterial species, instead of pairs of bacteria, which makes interpretation of results more meaningful.

$$M_i = \sum_{1\le j\le K} \hat{\theta}_j \cdot \log\left(X_j\right) \quad (5)$$

This final microbial signature is a log-contrast function, *i.e.* $\sum_{j=1}^{K} \hat{\theta} = 0$, which ensures the scale invariant CoDA principle. It also provides a convenient interpretation of the signature as a weighted

balance between two groups of bacteria, those with a positive coefficient vs those with a negative coefficient [9].

3. Results

To exemplify the application of the proposed methodology we used a dataset of simulated intestinal microbiome data from non-obese diabetic mice proposed by Koh, 2018 [10]. Samples from control mice at six weeks of age which were not exposed to any antibiotic treatment were used as template. Survival time, event (developing/not developing diabetes), censoring, age, and sex were also simulated in Koh's dataset. A final dataset of 100 samples and 353 different bacteria was used to perform the survival analysis using *coda4microbiome*. The aim of the analysis was to characterize a microbial signature able to predict the risk of developing diabetes.

The initial Cox model performed in the modelling step contained all possible pairwise log-ratios from the 353 bacteria's relative abundances, and it was adjusted by age and sex. Variable selection was performed with elastic-net and the optimal penalized parameter was stablished by cross-validation. It resulted in a model of 4 pairs of log-ratios with a mean cross-validation Harrell C index of 0.64 (±0.05). After reparameterization, the final microbial signature was expressed as a weighted balance between 4 bacterial species with positive coefficient and 3 with negative coefficient (Figure 1).

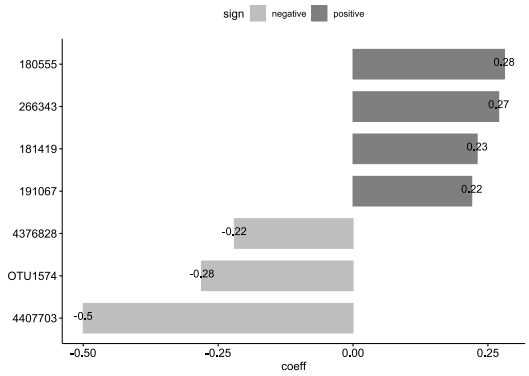


Figure 1: Bacterial species (vertical axis) with their respective coefficients that compose the microbial signature for the survival data analysis with *coda4microbiome.*

4. Conclusions

We introduced a new methodology for microbial analysis in survival studies that accounts for the compositional nature of microbiome data. The algorithm identifies a microbial signature that predicts the risk of a given event with the highest accuracy. Such signature is expressed as a balance between two groups of bacteria.

The algorithm has been implemented in R and it has been integrated in the existing R package *coda4microbiome* so that it can be easily used in survival studies to identify which microbial species are more associated to the development of a disease, response to a treatment, or even death.

5. Acknowledgments

This work was partially supported by the Spanish Ministry of Economy, Industry and Competitiveness, references PID2019-104830RB-I00 (M.L.C), PID2021-123657OB-C33 (A.S) and PID2021-122136OB-C21 (A.S.).

6. Bibliography

[1] Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genomics and Informatics*, *17*(1), e6. https://doi.org/10.5808/GI.2019.17.1.e6

[2] Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, *70*(1), 57–65.

[3] Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 1–18. https://doi.org/10.1186/s40168-017-0237-y

[4] Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., Jones, C. M. A., Wright, R. J., Dhanani, A. S., Comeau, A. M., & Langille, M. G. I. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, *13*(1), 1–16. https://doi.org/10.1038/s41467-022-28034-z

[5] Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society*, *44*(2), 139–177.

[6] Calle, M. L., & Susin, A. (2022). coda4microbiome: compositional data analysis for microbiome studies. *BioRxiv*, 2022.06.09.495511. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.06.09.495511v1

[7] Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220. http://www.jstor.org/stable/2985181

[8] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/JSS.V033.I01

[9] Susin, A., Wang, Y., Cao, K. A. L., & Luz Calle, M. (2020). Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics*, *2*(2). https://doi.org/10.1093/nargab/lqaa029

[10] Koh, H., Livanos, A. E., Blaser, M. J., & Li, H. (2018). A highly adaptive microbiome-based association test for survival traits. *BMC Genomics*, *19*(1), 1–13. https://doi.org/10.1186/s12864-018-4599-8

Contribution of blood DNA methylation to the association between smoking and lung cancer

Arce Domingo-Relloso,[1] Roby Joehanes,[2] Zulema Rodriguez-Hernandez,[3] Karin Haack,[4] M. Daniele Fallin,[5] Jason G. Umans,[6] Lyle G. Best,[7] Tianxiao Huan,[8] Chunyu Liu,[9] Jiantao Ma,[10] Jose D. Bermudez,[11] Shelley A. Cole,[12] Dorothy A. Rhoades,[13] Daniel Levy,[14] Ana Navas-Acien,[15] and Maria Tellez-Plaza[16]

[1] ad3531@cumc.columbia.edu. Department of Biostatistics, Columbia University, USA.

[2] roby.joehanes@nih.gov. National Heart, Lung, and Blood Institute, USA.

[3] zulema.rodriguez@isciii.es. National Center for Epidemiology, Spain.

[4] khaack@txbiomed.org. Population Health Program, Texas Biomedical Research Institute, TX, USA.

[5] dfallin@jhu.edu. University of Emory, GA, USA.

[6] jgu@georgetown.edu. MedStar Health Research Institute, Washington DC, USA.

[7] lbest@restel.com. Missouri Breaks Industries and Research Inc, SD, USA.

[8] Tianxiao.Huan@umassmed.edu. University of Massachusetts Medical School, MA, USA.

[9] liuc@bu.edu. Boston University School of Public Health, MA, USA.

[10] jiantao.ma@tufts.edu. Tufts University Friedman School of Nutrition Science and Policy, MA

[11] Jose.D.Bermudez@uv.es. Department of Statistics, University of Valencia, Spain.

[12] scole@txbiomed.org. Population Health Program, Texas Biomedical Research Institute, TX, USA.

[13] Dorothy-Rhoades@ouhsc.edu. Stephenson Cancer Center, USA.

[14] levyd@nhlbi.nih.gov. National Heart, Lung, and Blood Institute, USA.

[15] an2737@cumc.columbia.edu. Department of Environmental Health Sciences, Columbia University.

[16] m.tellez@isciii.es. National Center for Epidemiology, Spain.

Abstract

In this study, we investigated the potential mediating role of blood DNA methylation (DNAm) on the association between smoking and lung cancer. We extended the novel *multimediate* algorithm for multiple mediation analysis to survival data, which helped identify joint mediated effects of DNAm in the Strong Heart Study, with validation in the Framingham Heart Study. We additionally conducted functional validation using gene expression data, and bioinformatics analyses to confirm the biological plausibility of the findings.

Keywords: Smoking, DNA methylation, causal mediation analysis.

1. Introduction

Differential patterns in blood DNA methylation (DNAm) are associated with lung cancer, the main cause of cancer death worldwide,[1] suggesting that DNAm changes may play a key role in tumorigenesis.[2] Epigenetic signatures associated with smoking are robust across ethnically diverse populations,[3] and support that DNAm might be a causal intermediary in the biological pathway linking smoking to lung cancer.[4] However, studies investigating the role of DNAm in smoking-related lung cancer are unclear.[5] In most studies of smoking, DNAm and cancer are limited by the lack of time to

incident (i.e. newly diagnosed) cancer or the lack of formal mediation analyses. In addition, DNAm positions tend to be evaluated separately as independent mediators, however, given the high correlations between DNAm sites on the genome, considering them as joint mediators is more appropriate. To date, no algorithms have been developed that are able to conduct multiple mediation analysis with correlated mediators and survival outcomes.

In this study, we investigated whether the association of current and cumulative smoking with lung cancer risk might be explained by differences in human blood DNAm. We used data from the Strong Heart Study (SHS), a cohort of US Native Americans (discovery population), and the Framingham Heart Study (FHS) (replication population). We extended the *multimediate* algorithm,[6] which is able to conduct multiple mediation analysis in presence of correlated mediators, to a time-to-event setting, which enabled the evaluation of the most impactful DMPs potentially driving lung cancer risk. In addition, we conducted validation of the findings using whole blood gene expression in a subset of FHS participants, as well as a bioinformatic pathway enrichment analysis to assess the potential biological implication of the findings.

2. Methods

The SHS is a prospective cohort study of American Indian adults.[7] Blood DNAm was measured at baseline in 2,351 participants using Illumina's MethylationEPIC BeadChip. After preprocessing, 2235 individuals and 788,368 CpGs were included in this study. Lung cancer incidence was assessed by self-report during interviews, death certificates, chart reviews and pathology reports if available.

The FHS started in 1948. DNAm from whole blood was measured using the Illumina Infinium HumanMethylation 450K BeadChip. Gene expression from paired whole blood RNA was sequenced at >×30 depth of coverage using RNA-SeQC v1.1.9. according to TOPMed RNA-Seq pipeline v2. Cancer incidence was assessed by interviews, death certificates, and/or chart reviews that included pathology reports, and crosschecked with official medical records whenever possible.

Statistical methods

We first conducted a screening among the CpG sites that were associated with smoking in previous work in the SHS (303 CpGs in total),[8] by using a Cox ISIS coupled with elastic-net (ISIS-ENET, as conducted by the *SIS* R package), to select CpG sites associated with time to lung cancer. Models were adjusted by age, sex, BMI, study center, cell counts (CD8T, CD4T, NK, B cells and monocytes) and five genetic PCs. We calculated natural direct, indirect and total effects based on the product of coefficients method for mediation analysis using additive hazards models.[9] Mediated effects were reported as differences in cancer cases for current vs never smokers, or differences in cancer cases per a 10 cigarette pack-years increase, attributable to blood DNAm per 100,000 person-years.

We conducted functional validation of the genes identified in the mediation analysis by doing an expression quantitative trait methylation analysis (eQTM). We fitted a linear model for CpGs that were significant in the simple mediation analysis both in the SHS and the FHS. Batch effect-corrected expression was the dependent variable, batch effect-corrected DNAm was the predictor, and the model was adjusted for sex, age, predicted blood cell fraction, five expression PCs and 10 DNAm PCs, which accounted for population. We also conducted a KEGG enrichment analysis out of the genes annotated to cis- and trans- eQTMs to explore possible biological implications of our findings. The Kappa statistic, which is used to define KEGG terms interrelations (edges) and functional groups based on shared genes between terms, was set to 0.4. The enrichment analysis was performed using Cytoscape (version.3.8.2).

In presence of correlated mediators, traditional mediation analysis methods might lead to individual relative mediated effects that add up to more than 100 %, which suggests that some pathways

are overlapping and the joint and individual effects remain unidentifiable. To address this limitation, we extended the *multimediate* algorithm,[6] which uses the counterfactual multiple mediation framework, to the survival data setting using additive hazards models. The R code is available in Github (https://github.com/AllanJe/multimediate). This algorithm[6] is able to identify individual mediated effects of several mediators simultaneously while taking into account correlations between mediators.

Oncogenic transformations can happen several years before cancer diagnosis. Thus, as an attempt to discard cases where DNAm may have been measured after oncogenic transformations started, we repeated the mediation analysis excluding individuals with cancer that was diagnosed in the first 5 follow-up years (10 lung cancer cases excluded).

3. Results

The ISIS model selected 62 Differentially Methylated Positions (DMPs) associated with lung cancer. Of those, 29 DMPs had statistically significant indirect effects in the SHS for current versus never smoking. Among those, 20 were also measured in the FHS, of which 14 were replicated in the FHS. For cumulative smoking, 20 CpGs had statistically significant indirect effects in the SHS. Among those, 14 were also measured in the FHS, of which four were replicated in the FHS. The mediation models excluding cancer cases diagnosed during the first 5 follow-up years yielded similar results.

In the eQTM analysis, at a statistical significance p-value < 10^{-4}, 17 mediating DMPs of lung cancer in common for the SHS and FHS were associated with 12 cis-eQTMs and 2415 trans-eQTMs. The large majority of the eQTM-associated transcripts (75.7 % of transcripts in trans and 83.3 % of transcripts in cis) showed gene expression downregulation. Biological pathway enrichment analysis of target genes annotated to eQTM-associated transcripts showed 54 enriched biological pathways. Figure 1 displays overlapping DMPs, eQTMs and KEGG biological pathways by the evaluated exposures and endpoints. The enriched pathways were largely related to cancer.

In multi-mediator models, in absolute terms, of 385.7 (95% CI 265.9, 509.8) incident lung cancer cases per 100,000 person-years attributable to current smoking, 223.6 (95 % 126.1, 324.5), 62.6 (95 % CI 16.8, 110.2) and 28.3 (95 % CI 11.5, 46.5) lung cancer cases were attributable to differences in DNAm in cg05575921 (*AHRR*), cg24859433 (*IER3*) and cg11902777 (*AHRR*), respectively. This corresponds to 81.3 % of the effect of smoking in lung cancer driven by DNA methylation changes.

4. Discussion

We conducted a formal mediation analysis using time-to-newly diagnosed cancer data, and found that a substantial extent of the prospective association of smoking with lung cancer was explained by differences in blood DNAm. Results were largely consistent in the FHS, including additional validation of findings with expression data, which mostly showed methylation-related downregulation of distant genes that have a plausible role on cancer biological pathways. In the multimediator model, a joint mediated effect of 81.3 % was driven by three DMPs (annotated to *AHRR* and *IER3*).

Of note, our novel *multimediate* algorithm enabled us to explore the joint mediated effects of DMPs. Although many DMPs showed individual mediated effects in the single mediation analysis, the *multimediate* algorithm identified that the mediated effect was only driven by three DMPs. This means that many DMPs were identified as mediators by the single mediation analysis just because of having high correlations with actual mediators, but when considering them jointly in the same model, their contribution to the mediated effect was not significant. This fact highlights the importance of considering a multiple mediation approach as opposed to a simple mediation one.

This study has several limitations. First, although the replication in the FHS was high for lung cancer in the current versus never smoking model, it was smaller for lung cancer in the cumulative smoking model. Differences in smoking intensity and cessation across the SHS and FHS could explain some of the non-replicated DMPs. Also, non-fatal cancer data might be incomplete in the SHS as no linkage with the cancer registry is available. Despite these limitations, however, we still found substantial replication of findings between the SHS and the FHS.

Second, mediation analysis provides valid estimates only if the mediation assumptions such as absence of unmeasured confounding, which cannot be fully verified in practice, hold.[10] In addition, the *multimediate* algorithm is only valid in settings of non-causal correlations.[6] Experimental studies are needed to confirm the role of the identified blood DNAm signature of smoking in the association between smoking and smoking-related cancers.

Strengths of our study include replication in an independent cohort, the large sample size with methylation data from one of the largest microarrays nowadays available, the availability of information to account for numerous potential confounders and the additional validation of the results using gene expression data. In addition, we used state-of-the-art statistical methods including the novel *multimediate* algorithm for time-to-event data, which enabled the evaluation of correlated methylation sites jointly.

In conclusion, the prospective association of smoking with lung cancer in this study was largely explained by differences in few specific blood DNAm. These findings contribute to the identification of potentially novel mechanisms of lung cancer, and provide evidence in favor of DNAm as a potential biological intermediary in the association between smoking and lung cancer.

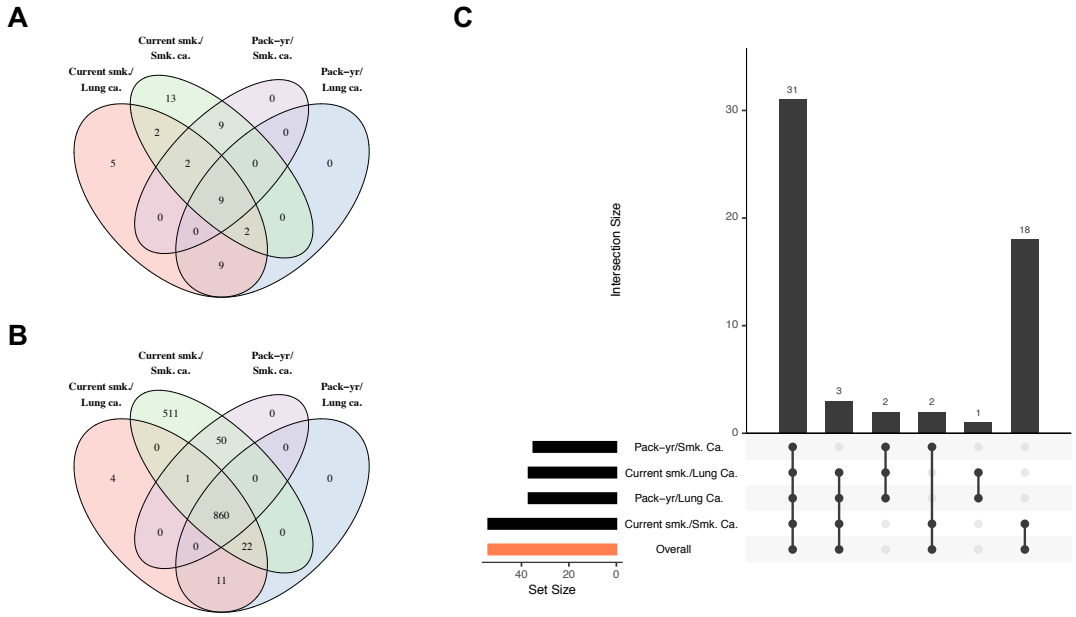


Figure 1: A) Venn diagram of CpGs with significant mediated effects both in the SHS and FHS. B) Venn diagram of genes annotated to the differentially expressed transcripts in trans in the Framingham Heart Study. C) Upset plot of the overlapping enriched KEGG pathways.

5. Bibliography

1. Bjaanæs MM, Fleischer T, Halvorsen AR, et al. Genome-wide DNA methylation analyses in lung adenocarcinomas: Association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis. *Mol Oncol*. 2016;10(2):330-343.
2. Klutstein M, Nejman D, Greenfield R, Cedar H. DNA Methylation in Cancer and Aging. *Cancer Res*. 2016;76(12):3446-3450.
3. Joehanes R, Just AC, Marioni RE, et al. Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet*. 2016;9(5):436-447.
4. Fasanelli F, Baglietto L, Ponzi E, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun*. 2015;6(1):10192.
5. Herceg Z, Ambatipudi S. Smoking-associated DNA methylation changes: No smoke without fire. *Epigenomics*. 2019;11(10):1117-1119.
6. Jérolon A, Baglietto L, Birmelé E, Alarcon F, Perduca V. Causal mediation analysis in presence of multiple mediators uncausally related. *Int J Biostat*. October 2020.
7. Lee ET, Welty TK, Fabsitz R, et al. The Strong Heart Study. A study of cardiovascular disease in American Indians: design and methods. *Am J Epidemiol*. 1990;132(6):1141-1155.
8. Domingo-Relloso A, Riffo-Campos AL, Haack K, et al. Cadmium, Smoking, and Human Blood DNA Methylation Profiles in Adults from the Strong Heart Study. *Environ Health Perspect*. 2020;128(6):067005.
9. Lange T, Hansen J V. Direct and Indirect Effects in a Survival Context. *Epidemiology*. 2011;22(4):575-581.
10. Zhang Z, Zheng C, Kim C, Van Poucke S, Lin S, Lan P. Causal mediation analysis in the context of clinical research. *Ann Transl Med*. 2016;4(21):425.

Generalized linear models with interval-censored covariates

Andrea Toloba[1,2], *Klaus Langohr*[2], *Guadalupe Gómez Melis*[2]

[1]andrea.toloba@upc.edu

[2]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya

Abstract

Interval-censored covariates have been appearing in clinical studies lately, but the lack of methodological content has not allowed researchers to analyse such data properly. We introduce the GEL technique for generalized linear models, and develop novel work on model checking. The proposed methodology is illustrated with data from metabolomics research area.

Keywords: interval censoring; generalized linear models; residual analysis

1. Introduction

Interval-censoring is typically encountered in the analysis of time-to-event data, and so statistical methods to analyse such data have been extensively studied; see, for example, Gómez et al (2009) for a thorough review [1]. On the contrary, scientific literature on regression models with an interval-censored covariate is rather scarce.

The GEL technique was first introduced in Gómez et al (2003) to estimate the regression parameters of a simple linear model where the independent variable is an interval-censored time [2]. Langohr et al (2014) revisited it and provided an implementation in `R` of the algorithm [3]. More recently, Morrison et al (2022) adapted the GEL technique to accommodate left-truncated interval-censored data, and Gómez et al (2022) rewrote the algorithm for generalized linear models [4, 5].

The definition of residuals has been even less explored than model estimation itself. Concerning linear regression, Langohr et al (2014) offer an exhaustive discussion of residuals in presence of interval-censoring [3]. The aim of this work is hence to propose a residual analysis procedure that accounts for interval-censoring, and address its suitability to check the model's goodness of fit.

2. Motivating data

The development of the methodology presented is motivated by metabolic data from 104 female participants of the PREDIMED-Plus trial. Clinical interest was on association of plasma carotenoid concentrations and cardiovascular risk factors [5]. Briefly, carotenoids are a family of eight phytochemical compounds produced by plants that are thought to be responsible for the health benefits fruits and vegetables provide, and are present in human blood acquired through diet. In order to study the global action of carotenoids as a whole, the sum of plasma concentrations was a relevant biomarker.

The plasma concentration of a carotenoid is measured in the laboratory by a technique called high-performance liquid chromatography. This technique is able to identify, extract and quantify the molecules of the compound, but it fails when concentrations are too low, so a mass spectrometry detection method is usually employed to obtain narrower intervals for the measure. Consequently, the data of a single plasma carotenoid determination depends on a limit of detection (LoD) and a limit of quantification (LoQ), in such a way that the value is exactly known over the LoQ, interval-censored between LoQ and LoD, and left-censored below LoD. These limits are compound-specific, so the sum of plasma concentrations gives rise to an interval-censored variable where the limits appear blurred, hence the resulting observed intervals are overlapping. This feature is an advantage over single determinations, for which very little information can be drawn under the limit of quantification.

3. Model formulation

For each carotenoid compound C_j, let C_{Lj}, C_{Rj} be two random variables denoting the observable intervals $[C_{Lj}, C_{Rj}]$, which correspond to the potential observations $[0, \text{LoD}_j)$, $[\text{LoD}_j, \text{LoQ}_j)$, and $[C_j, C_j]$. Note the observable intervals can be assumed closed because any quantification method has a limited decimal precision. The sum of carotenoids is denoted by Z, and the corresponding observable random variables are defined by $Z_L = \sum_{j=1}^{8} C_{Lj}$ and $Z_R = \sum_{j=1}^{8} C_{Rj}$. Hence, Z is an interval-censored variable, and the censoring interval $[Z_L, Z_R]$ verifies the non-informative conditions particular of this type of censoring. Figure 1 illustrates the observed data for one of the carotenoids, and for the sum of all of them. Additionally, let Y be a random variable from the exponential family denoting the response variable, and X a p-dimensional vector of fully-observed covariates.

Generalized linear models are an extension of classical linear models that aim to fit regression models with response variables whose distribution cannot be approximated by a normal random variable. These models relate the expected mean response $\mu = E[Y \mid X, Z]$ with the linear predictor $\eta = \alpha + \beta' X + \gamma Z$ through a monotonic differentiable link function g, that is $g(\mu) = \eta$. Many of the properties they possess arise from assuming that the probability density function of Y, in terms of η, pertains to the θ-parameter exponential family with shape

$$f(y; \theta, \phi) = \exp\Big\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y; \phi) \Big\},$$

where ϕ is known as dispersion parameter, and θ is a function of the mean $\mu = g^{-1}(\eta)$. For instance, an expression can be found for the variance of Y, $\text{Var}(Y) = a(\phi)\ddot{b}(\theta)$, which implies that it is not constant with respect to the mean μ; and the variance function $V(\mu)$ is defined so that $\text{Var}(Y) = a(\phi)V(\mu)$.

4. Parameter estimation

The GEL (Gómez - Espinal - Lagakos) technique is a parameter estimation method based on maximizing the log-likelihood function through an EM-type algorithm that contemplates an interval-censored covariate. It assumes the interval-censored covariate Z is discrete with support $\mathcal{S} = \{s_1, s_2, \ldots, s_m\}$ and probability law $w_j := P(Z = s_j)$ for $j = 1, \ldots, m$, yet no additional distributional restrictions are made.

Non-informative conditions are assumed in the sense that Z_L, Z_R provide no additional informa-

tion about Z than being inside the interval, and that Y depends on Z_L, Z_R only through Z. These are standard requirements for interval-censored data, and can be found for instance in Gómez et al (2009) [1].

Given n independent realizations of (Y, X, Z_L, Z_R), denote by $(y_i, x_i, z_{Li}, z_{Ri})$ the observed data of the ith individual. Non-informative conditions make it possible to ignore the distribution function of Z_L, Z_R from the log-likelihood maximization, since the likelihood function becomes proportional to

$$L(\theta, \phi, w) \propto \prod_{i=1}^{n} \sum_{j=1}^{m} \kappa_{ij}\ f(y_i \mid Z = s_j, X = x_i;\ \theta, \phi)\ w_j, \tag{1}$$

where $\kappa_{ij} = 1\{s_j \in [z_{Li}, z_{Ri}]\}$ indicates whether the support point s_j is included in the observed interval of the ith individual, and $w = (w_1, \ldots, w_m)$ is a parameter vector enclosing the probabilities of Z in $\mathcal{S}$. In order to estimate α, β, γ in presence of the nuisance parameter w, an iterative two-step algorithm is proposed:

1. Consider $\theta = \theta(\mu_i)$ fixed, and solve the following self-consistent equations for w.

$$w_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\kappa_{ij} f(y_i \mid Z = s_j, X = x_i;\ \theta) w_j}{\sum_{k=1}^{m} \kappa_{ik} f(y_i \mid Z = s_k, X = x_i;\ \theta) w_k}, \quad j = 1, \ldots, m$$

2. Given an estimate for w, the log-likelihood is maximized with respect to regression parameters α, β, γ. Notice that (1) does not correspond to the particular likelihood of generalized linear models, so standard tools such as iterative reweighted least squares or Fisher scoring algorithm are not applicable, and generic numerical algorithms for maximization shall be used instead.

5. Residual analysis

After model fitting it is indispensable to check the adequacy of the selected model to the data, since misspecification may invalidate findings and predictions. Common sources of model misspecification include a bad choice of link function g, a wrong assumption of the mean-variance relationship induced by the variance function $V(\mu)$, a wrong pre-specified dispersion ϕ, the possible omission of non-linear effects from covariates, and the presence of unusual response observations that might influence parameter estimation.

Diagnostic plots of residuals can reveal and help to identify these problems. For instance, plotting residuals against fitted values to seek patterns provides a first glimpse of systematic inconsistencies in the model, whereas non-linear effects can be checked with component-plus-residual plots, which show numeric covariates against its partial residuals. In addition, goodness-of-fit tests computed from residuals offer a valuable overview of model suitability.

It is well known that Pearson and deviance residuals, together with their standardized form, are the most employed in generalized linear models. Briefly, the former are defined as $r_{P,i} = (y_i - \hat{\mu}_i)/\sqrt{V(\hat{\mu}_i)}$, and arise from the Pearson's χ^2 statistic. Conversely, deviance residuals $r_{d,i} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$ are defined so that $D = \sum_i d_i$, where $D = 2\phi\{l(y) - l(\hat{\mu})\}$ is the deviance of the model. Unfortunately, both residuals rely on exact values of covariates, so are not totally suited for interval-censored data.

In this work we propose a residual analysis procedure based on the *pseudo-observations* estimating method originally conceived for incomplete data [6]. It will be compared with the alternative approach of computing the expected residuals $\hat{r}_{P,i} = E_Z[r_{P,i} \mid z_{Li}, z_{Ri}]$, $\hat{r}_{d,i} = E_Z[r_{d,i} \mid z_{Li}, z_{Ri}]$ under the Turnbull's non-parametric estimator of $P(Z = s_j)$.

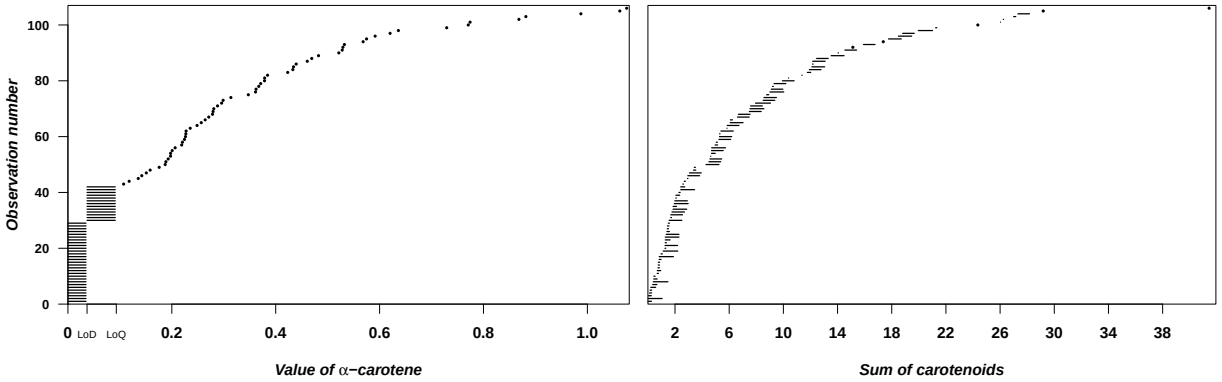


Figure 1: Ordered observed data from 104 female participants of a study on the association of circulating carotenoids and cardiovascular risk factors. Lines represent interval-censored observations, points exactly quantified determinations.

6. Aknowledgements

This work was funded by the Ministerio de Ciencia e Innovación (Spain) [Grant: PID2019-104830RB-I00/ DOI (AEI): 10.13039/501100011033] and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya (Spain) [Grant: 2021 SGR 01421].

7. Bibliography

[1] Gómez G., Calle M.L., Oller R., and Langohr K. (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling*, 9 (4), 259–297.

[2] Gómez G., Espinal A., and Lagakos S.W. (2003). Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine*, 22 (3), 409–425.

[3] Langohr K. and Gómez G. (2014). Estimation and residual analysis with R for a linear regression model with an interval-censored covariate. *Biometrical Journal*, 56 (5), 867–885.

[4] Morrison D., Laeyendecker O., and Brookmeyer R. (2022). Regression with interval-censored covariates: Application to cross-sectional incidence estimation. *Biometrics*, 78 (3), 908–921.

[5] Gómez G., Marhuenda-Muñoz M., and Langohr K. (2022). Regression analysis with interval-censored covariates. Application to liquid chromatography. In Sun J., Chen DG. (eds) *Emerging topics in modeling interval-censored survival data*. ICSA Book Series in Statistics. Springer, Cham.

[6] Andersen P.K., Klein J.P., and Rosthøj S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90 (1), 15–27.

Sesión Especial:
Italia – Portugal

AUC estimation in logistic regression with missing data: a case study

Susana Rafaela Martins[1], *Jacobo de Una-Alvarez*[2], *Maria del Carmen Iglesias-Perez* [3]

[1]srgm@estg.ipvc.pt, Escola Superior de Desporto e Lazer, Instituto Politecnico de Viana do Castelo & SiDOR research group, Universidade de Vigo

[2] jacobo@uvigo.es, Department of Statistics and OR & CINBIO, Universidade de Vigo

[3] mcigles@uvigo.es, Department of Statistics and OR & CINBIO, Universidade de Vigo

The obesity is defined by the excessive accumulation of fat and the consequent excess weight, which translate into a risk to the individual's health. According to the WHO, the levels of overweight and obesity continue to increase, particularly in young people and children between 5 and 19 years of age. Globally, in 2016 the prevalence levels of overweight and obesity were around 18%. This problem has a solution, however its prevention is the best option. In this sense, and taking advantage of data from a previously existing study, we decided to study the possibility of building predictive models that allow monitoring obesity levels. The aforementioned study was carried out with children from the municipality of Viana do Castelo, in Portugal. The initial data was collected in 1997, and annually until the year 2000, with a subsequent collection in 2007. Naturally, one of the problems that arise in this type of study is the existence of missing data.

In this work that we present, we try to monitor the levels of childhood obesity and overweight using the logistic model to predict them through covariates related to performance in physical test in addition to gender and previous levels of overweight. The importance of this study is related to the fact that the results of the physical tests can be easily collected by any sports teacher in a school context and, consequently, it can be easy to screen children who may be at risk of having overweight or obesity. To study the predictive capacity of the logistic model, the area under the ROC curve (AUC) was used.

In this work we investigate the issue of estimating the AUC in presence of missing data. A simulation study is carried out to compare the performance of several approaches: Complete Case Analysis, Inverse Probability Weighting and Multiple Imputation. We also take into account the problem of the optimistic estimation of the AUC.

Keywords: Obesity, prediction, ROC curve

Modelling the hazard of transition into the absorbing state in the illness-death model

Elena Tassistro[1], *Davide Paolo Bernasconi*[2], *Paola Rebora*[3], *Maria Grazia Valsecchi*[4], *Laura Antolini*[5]

[1]elena.tassistro@unimib.it, Bicocca Center of Bioinformatics, Biostatistics and Bioimaging (B4 Centre), School of Medicine and Surgery, University of Milano-Bicocca

[2]davide.bernasconi@unimib.it, Bicocca Center of Bioinformatics, Biostatistics and Bioimaging (B4 Centre), School of Medicine and Surgery, University of Milano-Bicocca

[3]paola.rebora@unimib.it, Bicocca Center of Bioinformatics, Biostatistics and Bioimaging (B4 Centre), School of Medicine and Surgery, University of Milano-Bicocca

[4]grazia.valsecchi@unimib.it, Bicocca Center of Bioinformatics, Biostatistics and Bioimaging (B4 Centre), School of Medicine and Surgery, University of Milano-Bicocca

[5]laura.antolini@unimib.it, Bicocca Center of Bioinformatics, Biostatistics and Bioimaging (B4 Centre), School of Medicine and Surgery, University of Milano-Bicocca

The illness-death model is the simplest multistate model where the transition from an initial state 0 to an absorbing state 2 may involve also an intermediate state 1. The impact of the transition into 1 on the subsequent transition hazard to 2 enables to increase the knowledge about the disease evolution. The standard analysis approach is modelling the transition hazards from 0 to 2 and from 1 to 2 including time to illness as a time-varying covariate and measuring time from origin even after the transition into 1. The hazard from 1 to 2 can be also modelled only on patients in state 1, measuring time from illness and including time to illness as a fixed covariate. A recently proposed approach is a model where time after the transition into 1 is measured in both scales and time to illness is included as a time-varying covariate. Another possibility is a model where time after the transition into 1 is measured only from illness and time to illness is included as a fixed covariate. This work aims to set up a strategy a statistician can follow to fit the most suitable full-sample model on the hazards of transition to state 2.

Through theoretical reasoning and simulation protocols we developed sequential strategies a statistician can follow to: a) validate the properties of the illness-death process, from which the choice of the scale to measure time after illness depends, b) estimate the impact of time to illness on the hazard from 1 to 2, proposing also a novel modelling approach that ensures the interpretability of the coefficient of the time to illness.

In the case of Markov data, the use of the clock forward time scale is the natural way to measure the follow-up time. The clock reset scale should be considered in case of non-Markov data, since forcing to use the clock forward scale will result in a spurious effect of the time to illness, due to the time after illness and not to a different shape of the hazard function after illness.

Keywords: illness-death, Markov model, time scales.

A Practical Guide to Analyzing Complex Survival Data with Kaplan-Meier

Luís Meira-Machado[1]

[1]lmachado@math.uminho.pt, Department of Mathematics and Centre of Mathematics, University of Minho, Portugal

The estimation of survival has been a widely researched topic in statistical and medical literature. Among the commonly used estimators, the Kaplan-Meier product-limit estimator is preferred due to its nonparametric nature, which does not rely on any assumptions about the lifetime probability distribution. The estimator is computed using a product of elementary probabilities, which calculates conditional survival probabilities. The redistribution to the right algorithm is another method that is employed to estimate the Kaplan-Meier estimator of survival by redistributing the mass of a censored subject equally among those who fail or are censored at later times. In this work, we present additional alternative representations of the Kaplan-Meier estimator and discuss their applications and advantages of its usage. One of these representations defines the estimator as a sum of weights, which is useful in estimating various quantities in the context of multi-state models. The estimator can also be represented as a weighted average of identically distributed terms, where the weights are obtained by using the inverse probability of censoring. We also demonstrate how these formulations can be applied to estimate different quantities, particularly in the context of multi-state models. Additionally, we will also focus on methods that aim to reduce the variability of the estimators and allow for the estimation of some of these quantities conditionally on covariates. To illustrate these methods, we include two real data examples from medicine.

Keywords: Kaplan-Meier, Multi-state models, Survival Analysis

Adverse events with survival outcomes: from clinical questions to methods for statistical analysis

Tassistro Elena[1], *Valsecchi Maria Grazia*[2], *Bernasconi Davide Paolo*[3], *Antolini Laura*[4]

[1]elena.tassistro@unimib.it, School of Medicine and Surgery, University Milano Bicocca, Italy

[2]grazia.valsecchi@unimib.it, School of Medicine and Surgery, University Milano Bicocca, Italy

[3]davide.bernasconi@unimib.it, School of Medicine and Surgery, University Milano Bicocca, Italy

[4]laura.antolini@unimib.it, School of Medicine and Surgery, University Milano Bicocca, Italy

Keywords: adverse events, competing risks.

Introduction and Objective(s) - When studying a novel treatment with a survival time outcome, failure can be defined to include an adverse event (AE) among the endpoints typically considered, for instance relapse. These events act as competing risks, where the occurrence of relapse as first event and the subsequent treatment change exclude the possibility of observing AE related to the treatment itself. In principle, the analysis of AE could be tackled by two different approaches: 1. It requires a competing risk framework for analysis: the clinical question relates to the observed occurrence of AE as first event, in the presence of the event "relapse"; 2. It requires a counterfactual framework for analysis: the clinical question relates to the treatment causing AE occurrence as if relapse could not occur. This work has two aims: the first is to critically review the standard theoretical quantities and estimators with reference to their appropriateness for dealing with approaches 1 or 2 and to the following features: (a) estimators should address for the presence of right censoring; (b) theoretical quantities and estimators should be functions of time. The second aim is to define a strategy to relax the assumption of independence between the potential times to the competing events of the commonly used estimators when counterfactual approach 2 is of interest.

Method(s) and Results - After reviewing the standard methods[1] we clarify the impact of the crucial assumption of independence between potential times to competing events of the standard estimators used in the counterfactual approach. We propose the use of regression models, stratified Kaplan-Meier curves and inverse probability of censoring weighting[2] to relax the assumption of independence by achieving conditional independence given covariates and we develop a simulation protocol to show the performance of the proposed methods.

Conclusions - The proposed methods overcome the problem due to the dependence between the two potential times. In particular, one can handle patients' selection in the risk sets, and thus obtain conditional independence between the two potential times, adjusting for all the observed covariates that induce dependence. The proposed methods can be also extended to the case of repeated adverse events.

References

[1] A. Allignol, J. Beyersmann, C. Schmoor (2016). Statistical issues in the analysis of adverse events in time-to-event data. Pharmaceutical Statistics, 15, 297-305

[2] S.J.W. Willems, A. Schat, M.S. van Noorden, M. Fiocco (2018). Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. Biometrical Journal, 62, 836-851

Sesión Especial:
Sociedad Española de Epidemiología

Statistical considerations for analysing data derived from longitudinal cohort studies

Rocío Fernández-Iglesias[1], *Pablo Martinez-Camblor*[2], *Ana Fernández-Somoano*[3]

[1]rocio.fdez.iglesias@gmail.com; Spanish Consortium for Research on Epidemiology and Public Health (CIBERESP), Madrid, Spain; University Institute of Oncology of the Principality of Asturias (IUOPA), University of Oviedo, Asturias, Spain; Health Research Institute of the Principality of Asturias (ISPA), Asturias, Spain.

[2]pablo.martinez-camblor@hitchcock.org; Biomedical Data Science Department, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA; Faculty of Health Sciences, Universidad Autónoma de Chile, Providencia, Chile.

[3]fernandezsana@uniovi.es; Spanish Consortium for Research on Epidemiology and Public Health (CIBERESP), Madrid, Spain; University Institute of Oncology of the Principality of Asturias (IUOPA), University of Oviedo, Asturias, Spain; Health Research Institute of the Principality of Asturias (ISPA), Asturias, Spain.

Modern science is frequently based on the exploitation of large volumes of information storage in data sets, and involving complex computational architectures. The statistical analyses of these datasets have to cope with specific challenges, and frequently involve making more or less arbitrary decisions. Epidemiological papers have to be concise and use to be focused on the underlying clinical or epidemiological results, not paying too much attention to some relevant methodological details. In this work, we use an analysis of the cardiovascular-related measures tracking in 4-8 years children using data from the INMA-Asturias project for illustrating how the making-decision process was performed. We focus on two particular aspects of the problem, how to deal with missing data and which regression model to apply to evaluate tracking when there are no defined thresholds to categorize variables into risk groups. As spoiler, we discuss what were the results of our multiple imputation analyses and which were the advantage of using quantile regression models to analyse the tracking.

Keywords: Missing data; quantile regression, tracking.

Reconstruction of smoking prevalence in Spain by sex and age groups in the period 1991-2020

Guerra-Tort C[1], *López-Vizcaíno E* [2], *Santiago-Pérez MI*[3], *Rey-Brandariz J*[1], *Varela L*[1], *Candal C*[1], *Ruano-Ravina A*[1], *Pérez-Ríos M*[1]

[1]carla.guerra@rai.usc.es, Departamento de Medicina Preventiva e Saúde Pública, Universidade de Santiago de Compostela

[2]esther.lopez@ige.eu, Servizo de Difusión e Información, Instituto Galego de Estatística, Xunta de Galicia

[3] Servizo de Epidemioloxía, Dirección Xeral de Saúde Pública, Xunta de Galicia

In Spain, there is incomplete available information to make a global assessment of the evolution of the smoking epidemic. In this work we propose to accurately reconstruct the annual series of smoking prevalence in our country in the period 1991-2020 applying a small-area model.

The data used are derived from public statistics, with special relevance to those derived from the different National (1993, 1995, 1997, 2001, 2003, 2006, 2011, 2017) and European (2009, 2014, 2020) health surveys. A small-area estimation method based on aggregated data was used to reconstruct smoking prevalence by sex and five-year age group (from 15-19 years to 80-84 years) for each year of the period 1991-2020. Specifically, a small-area multinomial logistic mixed model with area and time random effects was applied. The areas were the D=30 groups defined from crossing the 15 five-year age groups with sex, and the time periods were the T=30 years of the 1991-2020 series. In the model, the response variable is a vector with the number of smokers, ex-smokers and never smokers in each area and time, and the covariates are aggregated information related to tobacco consumption obtained from registries (sociodemographic, economic and morbidity data). The model is expressed as:

$$p_{dkt} = \frac{exp\ (\eta_{dkt})}{1 + exp(\eta_{d1t}) + exp\ (\eta_{d2t})},$$

$$\eta_{dkt} = \log\left(\frac{p_{dkt}}{p_{d3t}}\right) = x_{dkt}\beta_k + u_{1,dk} + u_{2,dkt},\ \ d = 1, \dots, D, k = 1,2, t = 1, \dots, T,$$

where p_{dkt} is the prevalence of each category k corresponding to area d and time t, $x_{dkt} = (x_{dkt1}, \dots, x_{dktr_k})'$ is the set of covariates corresponding to category k, area d and time t, and $\beta_k = (\beta_{k1}, \dots, \beta_{kr_k})'$ is the vector of regression parameters. The subscript k refers to the category of smokers (k=1) or ex-smokers (k=2). The third category of never smokers (k=3) is taken as the reference category. The model also considers random effects $u_{1,dk}$ and $u_{2,dkt}$ associated with area d and category k, and area d and category k and time t, respectively. To fit the model, we combine the penalized quasi-likelihood method, for the estimation and prediction of β_{kr_k}, $u_{1,dk}$ and $u_{2,dkt}$, with the restricted maximum likelihood method, which is used to estimate variance components.

To estimate the prevalence of tobacco consumption the following steps were carried out:1) union of the 11 survey's data and computation of the individual smoking status of the people polled from current status, age at initiation and age at cessation for each year of the period 1991-2020; 2) calculation of the prevalence of smokers, ex-smokers and never smokers by sex, age group and year applying a weighted ratio estimator; 2) preparation of covariates by sex, age group and year; 3) selection of covariates to fit the model for smokers and ex-smokers, and 4) adjustment of the small-area model.

We expect that the results of this work and its methodology can be applied to other lifestyle and health determinants such as alcohol consumption or exposure to second-hand smoke, allowing us to quantify their detailed impact on the health of populations.

Keywords: small-area analysis, smoking, health surveys.

Association of plasma metabolomic compounds with the incidence of cardiovascular endpoints in the Hortega Follow-Up Study

[1,2]Zulema Rodriguez-Hernandez, [3]Pilar Casanovas, [1]Marta Galvez-Fernandez, [3,4]Vannina Gonzalez-Marrachelli, [5]Arce Domingo-Relloso, [3]Maria Grau-Perez, [6]Laisa Briongos-Figuero,[6]Juan C. Martin-Escudero, *[1]Maria Tellez-Plaza, *[3]Josep Redon, *[3,7]Daniel Monleon*

* Equal author contribution, [1]Integrative Epidemiology Group,National Center for Epidemiology, ISCII, Madrid, Spain; [2]Universitat Politècnica de València, Valencia; [3]INCLIVA Biomedical Research Institute, Valencia, Spain; [4]Department of Physiology, University of Valencia, Spain; [5]Department of Environmental Health Sciences, Columbia University, New York, US; [6]Department of Internal Medicine, Hospital Universitario Rio Hortega, Valladolid, Spain; [7]Department of Pathology, University of Valencia, Valencia, Spain

Introduction: The association of metabolic compounds with the incidence of specific cardiovascular (CV) endpoints including coronary heart disease (CHD), stroke and heart failure (HF) has rarely been studied in general population settings. Therefore, we evaluated the prospective association of metabolic compounds with incidence of CHD, stroke and HF in the Hortega study, a representative sample of a general population from Spain.

Methods: Metabolites were measured by NMR in 1016 adults of the Hortega Follow-up Study (15 years of follow-up) without clinical CV diseases at baseline. We estimated hazard ratios (HR) and 95% confidence intervals (CI) of stroke, CHD and HF incidence by plasma metabolites levels (log-transformed) using Cox proportional hazards regression. Models were adjusted for sex, education, smoking status, cumulative tobacco smoking (pack-year), urine cotinine, glomerular filtration rate, physical activity, HDL cholesterol, total cholesterol, lipid lowering and blood pressure medication, type 2- diabetes mellitus and systolic blood pressure.

Results: The number of newly diagnosed cases were 67 for stroke over 13,184 person-years (incidence of 5.1 per 1,000 person-years); 52 for CHD over 12,908.5 person-years (incidence of 4 per 1,000 person-years) and 75 for HF over 13,336.3 person-years (incidence of 5.6 per 1,000 person-years). We observed statistically significant associations [HR (95% CI), comparing the 80^{th} to the 20^{th} percentiles of metabolites distributions] for creatinine phosphate [1.94 (1.18, 3.20)], tryptophan [3.25 (1.48, 7.15)], tyrosine [1.98 (1.18, 3.32)] and O-phosphoethanolamine [2.09 (1.25, 3.50)], among others, with incident heart stroke; cysteine [2.12 (1.23, 3.68)], isopropanol [2.25 (1.20, 4.21)], citrate [2.20 (1.20, 4.01)] and phenylpropionate [2.45 (1.13, 5.31)], among others, with incident CHD; and some fatty acids subclasses as CH_2CH_2CO [0.56 (0.32, 0.97) and CH_2N [0.44 (0.22, 0.88)], acetone [0.45 (0.22, 0.93)] and lactate [0.53 (0.28, 0.99)] with incident HF.

Conclusions: Metabolic patters reflecting amino acids, fatty acids, microbiota co-metabolism and energy-related compounds were prospectively associated with specific CV endpoints in the general population from Spain, which may be relevant for CV diseases prevention and diagnosis. Additional studies for reproduction of our findings are needed.

Keywords: metabolomics, metabolites, cardiovascular disease

Bayesian-spatial distributed lag non-linear models: A temperature-mortality case study in Barcelona

Marcos Quijal-Zamorano[1,2]; Miguel A. Martinez-Beneito[3]; Joan Ballester[1]; Marc Marí-Dell'Olmo[4,5,6]

Affiliations:

1 ISGlobal, Barcelona, Spain

2 Universitat Pompeu Fabra (UPF), Barcelona, Spain

3 Departament d'Estadística i Investigaciò Operativa, Dr. Moliner 50, 46100, Burjassot, Valencia, Spain

4 Agència de Salut Pública de Barcelona (ASPB), Pl. Lesseps 1, 08023 Barcelona, Spain

5 Institut d'Investigació Biomèdica Sant Pau (IIB SANT PAU), Sant Quintí 77-79, 08041 Barcelona, Spain

6 Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Av. Monforte de Lemos, 3-5, 28029 Madrid, Spain

In the context of climate change and increasing temperatures, the interest in the health effects of environmental exposures has remarkably increased. The development of the distributed lag non-linear models (DLNM) has become rapidly the referent framework when studying temperature-mortality short-term associations. DLNMs facilitates the modelling of the non-linear and lagged effect of temperatures on mortality. However, the small-area analysis of temperature-mortality is still scarce. In that sense, here we present four models. The first two models generalize standard DLNMs to a Bayesian framework, using a case-crossover design (model 1) and the common DLNM time-series configuration, where time trend and seasonality are modelled by using splines (model 2). We propose models 3 and 4 specifically for dealing with unstable estimates from small numbers in small-area analyses. These two models are extensions of model 1 and model 2 respectively, where we use Leroux models to spatially-smooth in one-stage approaches the coefficients of the exposure-response relationships for each small area. We apply all proposed models to a case-study for assessing the temperature-mortality relationships in the 73 neighborhoods of Barcelona during summer months. 39.569 deaths were considered in the period 2007-2016, 19 of them corresponding to the neighborhood with the lowest number of deaths and 1.454 deaths to the one with the highest number. Curves defining the relative risks of mortality were unstable and unreliable in the independent models, with regions with extremely high and low risks distributed all over the city. Spatial models benefit from adjacent regions to smooth the association and reveal hidden spatial patterns of risk. In addition, the flexibility of these Bayesian models allowed us to explore the results of these epidemiological models in new intuitive ways. This novel multidimensional approach brings the opportunity to estimate ecological temperature-mortality models in a smaller spatial scale to better understand the socioeconomic and built environment factors driving the effect of temperature on human health.

Keywords: temperature, mortality, Bayesian, distributed lag non-linear models, spatial models, small-area.

Sesión Especial:
Red Nacional de Bioestadística
BIOSTATNET

BIOSTATNET: advancing in the research of excellence in biostatistics at national and international level

David Conesa[1], *on behalf of the whole Biostatnet Network*

[1]david.v.conesa@uv.es, Department of Statistics and Operations Research, University of Valencia

In 2010, BIOSTATNET was established as a thematic network with the goal of creating a competitive framework for cohesion and coordination of Biostatistics research, teaching, and transfer in Spain. Its mission is to serve as a central hub connecting Biostatistics researchers from various research projects within the state RDI plan.

Inspired by Carmen Cadarso, who acted as the PI of the network till she passed away in 2022, and firstly structured in 8 nodes coordinated by her and Susie Bayarri (later by Carmen Armero), Maria Durbán, Guadalupe Gómez, Jesús López-Fidalgo, Antonio Martín, Vicente Nuñez-Antón and Pere Puig, the network has been consolidating as a functional and efficient structure, and has allowed: a) to coordinate research and teaching in Biostatistics in Spain with international projection, b) to promote appropriate training in Biostatistics; and c) to promote its transfer and applicability in life and health sciences.

With 10 nodes and more than 250 researchers, the network is currently undergoing a period of change, with new nodes being created and new leaders on board. The aim is to continue the work that has been done and bring fresh blood into the network. Here, we will present the main characteristics of the network, including its nodes and research, the types of activities undertaken by the network, and the process for becoming a member.

Keywords: Biostatistics, Network, Research.

A joint modelling approach for Health-Related Quality of Life and survival analysis of a 5-year follow-up study of COPD patients

Cristina Galán-Arcicollar[1], *Josu Najera-Zuloaga*[2], *Dae-Jin Lee*[3], *Cristobal Esteban*[4], *Inmaculada Arostegui*[1,2]

[1]{cgalan,iarostegui}@bcamath.org, Applied Statistics Research Line, Basque Center for Applied Mathematics, Bizkaia, Spain

[2]{josu.najera,inmaculada.arostegui}@ehu.eus, Department of Mathematics, University of the Basque Country UPV/EHU, Bizkaia, Spain

[3] daejin.lee@ie.edu, School of Science and Technology, IE University, Madrid, Spain

[4] cristobal.estebangonzalez@osakidetza.eus, Servicio de Neumología, Hospital Galdakao-Usansolo, Galdakao, Bizkaia, Spain

Recently, in clinical trials, there has been an increasing interest in using longitudinal biomarkers for characterizing the occurrence of an event, such as death or illness recovery. This interest leads to longitudinal studies based on patients' follow-up periods where the values of certain variables may be recorded repeatedly, whereas the time-to-event is also monitored. Therefore, two types of outcomes from the same subject are simultaneously observed: repeated measures and time-to-event. The inherent association between the outcomes has brought the joint modelling framework to analyze them jointly.

Furthermore, there is a growing priority on placing patients at the centre of healthcare research and evaluating clinical care. In this context, patient-reported outcomes (PROs) are helpful tools for informing clinicians about patients' health status. This information comes directly from the patients and is collected by providing them with questionnaires that consider their health, quality of life, or functional status.

In this work, we propose a joint modelling approach for longitudinal PRO measurements and survival data that include adequate distributional fits of PRO by considering its nature and characteristics. In particular, we assessed data from a 5-year follow-up study of 543 patients with chronic obstructive pulmonary disease (COPD) from Galdakao-Usansolo Hospital in Biscay, Spain. The overall impact of COPD on the subject is multifaceted. Thus, more than clinical biomarkers are needed to assess the disease evolution. In this sense, the COPD study considered survival data and one to four Health-Related Quality of Live (HRQoL) scores per individual collected during the follow-up period. Two questionnaires were used to evaluate the HRQoL, which turned out to be an important indicator of the health status of patients with chronic diseases: one generic, the Short-Form 36, and the other disease-specific, St. George's Respiratory Questionnaire. We provide relevant clinical results over the association of the HRQoL and survival data in the COPD study. Additionally, we performed a simulation study based on the COPD study scenario to compare our proposal with two popular approaches that also consider both outcomes: the Joint Modelling full likelihood approach and the Time-Varying covariate Cox model.

Keywords: Joint Modelling, PRO, COPD.

Shared parameter models for the analysis of randomized clinical trials whose primary endpoint is a normally distributed longitudinal response

Alberto García-Hernandez[1], Teresa Pérez Pérez[2], María del Carmen Pardo Llorente[3], Dimitris Rizopoulos[4]

[1]albega28@ucm.es, Facultad de Estudios Estadísticos, Univ. Complutense, Madrid

[2]teperez@estad.ucm.es, Facultad de Estudios Estadísticos, Univ. Complutense, Madrid

[3]mcapardo@mat.ucm.es, Facultad de Ciencias Matemáticas, Univ. Complutense, Madrid

[4]d.rizopoulos@erasmusmc.nl, Erasmus Medical Center, Rotterdam

Introduction: In a randomized clinical trial, it is crucial to adhere to the intent-to-treat principle and handle appropriately missing data. An intercurrent event (ICE) refers to any circumstance that occurs during the study that affects the interpretation and/or observation of the response of interest, such as a participant's decision to discontinue the study medication. Every clinical trial should have a well-defined study estimand, that is, the specific treatment effect that is being estimated. The study estimand can only be well-characterized when all relevant ICEs are identified, and a strategy to handle each of them is pre-specified in the study protocol. It is the responsibility of the study statistician to choose an estimator that properly handles missing data in accordance with the chosen estimand.

Methodology: We examined the performance of shared parameter models (SPM) in analysing normally distributed longitudinal endpoints in the presence of an intercurrent event that needs to be handled with a hypothetical strategy. Using this strategy, we are not interested in using data collected after the ICE as our estimand targets an imaginary world where the ICE has not occurred. With SPM, we modelled the (longitudinal and normally distributed) endpoint of interest together with the time-to-event process associated with the occurrence of the ICE. We compared SPM with the current gold standard methodology in this field, mixed models for repeated measures (MMRM). Using MMRM, the probability distribution of the ICE is ignored as this process is simply considered an underlying generator of missingness. Additionally, we proposed a new methodology to choose between MMRM and SPM by expanding the longitudinal data density (using MMRM) into the likelihood of both longitudinal and time-to-event data by plugging in the likelihood of a survival parametric time-varying covariates model.

Results: The simulation study demonstrated that the SPM approach outperforms MMRM in terms of bias only if the association between the endpoint of interest and the ICE follows the SPM parameterization. However, SPM introduced significant bias when the ICE process depended not only on the random effects but also on the entire last observation (including noise) of the longitudinal response. Additionally, SPM was rather sensitive to the correct specification of the association structure. The simulation experiment also showed that the novel approach proposed to choose between MMRM and SPM accurately selects the optimal tool (MMRM or SPM) with sample sizes typical of phases 2b and 3.

Conclusions: This research has highlighted some limitations associated with the use of SPM for the analysis of longitudinal responses in randomized clinical trials. In particular, we have demonstrated that SPM outperforms standard mixed models only under a narrow set of conditions. However, we have proposed a novel methodology that can accurately select the optimal tool (MMRM or SPM) for a given set of conditions.

Keywords: missing data, estimands, shared parameter models.

A Bayesian competing risks survival model to study the cause of death in patients with heart failure

Jesús Gutiérrez-Botella[1], *Carmen Armero*[2], *María Pata*[3], *Thomas Kneib*[4], *Francisco Gude-Sampedro*[5]

[1]jesus.gutierrez.botella@rai.usc.es, Biostatech, Advice, Training and Innovation in Biostatistics SL; GRID-BDS, University of Santiago de Compostela
[2]carmen.armero@uv.es, Department of Statistics and OR, Universitat de València
[3]mariapata6@biostatech.com, Biostatech, Advice, Training and Innovation in Biostatistics SL
[4]tkneib@uni-goettingen.de, Georg-August-Universität Göttingen
[5]francisco.gude.sampedro@sergas.es, Epidemiology Department, Clinical University Hospital of Santiago de Compostela

Heart Failure (HF) is a chronic, progressive condition which happens when the heart is not able to pump enough blood to supply the patient's tissues. Cardiac Resynchronization Therapy (CRT) is a procedure to implant electrodes in the heart's chambers to make the heart work in a more organized and efficient way. This therapy improves the prognosis and reduces hospitalization rates and mortality in HF patients. Although the effects of this therapy have been assessed on a short-term basis, there are scarce published data on the long-term benefits of the CRT.

The aim of this work is to study the long-term cardiovascular and non-cardiovascular death in HF patients who underwent CRT and its relationship with demographic and clinical variables. This follow-up study includes 296 patients who received CRT in a tertiary cardiac institution between August 2001 and April 2015. Patients with unknown cause of death were withdrawn.

For the statistical analysis we used a Bayesian competing risks model for the events cardiovascular death and non-cardiovascular death. Bayesian estimation was performed using MCMC methods with JAGS software. Posterior outputs such as the posterior distribution for the cause-specific baseline hazard function for cardiovascular and non-cardivascular death, posterior distribution for the cumulative incidence function for each cause of death as well as the posterior distribution of the overall survival function are discussed.

Keywords: Cardiac Resynchronization Therapy; Cardiovascular and non-cardiovascular death; Cumulative incidence function; Overall survival function.

Comunicaciones Orales

Joint modelling of several diseases for high-dimensional spatial data using a multivariate scalable Bayesian approach

A. Adin[1,2], *T. Goicoa*[1,2], *G. Vicente*[3] *and M.D. Ugarte*[1,2]

aritz.adin@unavarra.es, tomas.goicoa@unavarra.es, gonzalo.vicente@fce.uncu.edu.ar, lola@unavarra.es

[1] Department of Statistics, Computer Science and Mathematics, Public University of Navarre, Spain
[2] Institute for Advanced Materials and Mathematics (INAMAT[2]), Public University of Navarre, Spain
[3] Facultad de Ciencias Económicas, Universidad de Cuyo, Argentina

Research on spatial multivariate disease mapping has received considerable attention in recent years, although the use of these models remains limited due to difficulties in their implementation and computational burden. These problems are accentuated when the number of small areas is very large. In this work, we introduce a scalable, order-free scalable Bayesian modelling approach to jointly smooth mortality or incidence risks of several diseases for high-dimensional areal count data. Our proposal partitions the spatial domain into smaller subregions, fits multivariate models in each subdivision and obtains posterior distribution of relative risks across the entire spatial domain. The modelling approach also provides local posterior estimates of between-disease correlations and variance parameters in each partition that are combined through a consensus Monte Carlo algorithm to obtain global estimates for the whole study region. We implement the proposal using integrated nested Laplace approximations (INLA) to reduce the computational burden through the `R` package `bigDM`, which also implements recent high-dimensional univariate scalable models for spatial and spatio-temporal count data.

Keywords: INLA, Non-stationary models, Spatial epidemiology.

Extreme events in the framework of species distribution models: a Bayesian approach

L. Aixalà[1], *X. Barber*[1], *D.V. Conesa*[2], *A. López-Quílez*[2]

[1]laixala@umh.es , Centro de Investigación Operativa. Universidad Miguel Hernández de Elche
[2] Departament d'Estadística i I.O. Universitat de València

Extreme events have been studied in the field of ecology for many years, either from a purely descriptive point of view in its principles, or last decades from an inferential and/or predictive perspective to try to understand them.

The first problem to be addressed in the study of extreme events is the definition of the event itself: can we speak of an extreme event with a simple 2% daily mortality of a species if its usual mortality is 0.5%, or can we speak of an extreme event if the sea temperature increases by $2°$C in 48 hours?

In the present work, included in a project to study extreme events and climate change, we want to study the relationship between certain types of extreme events (DANA, heat waves, etc.) and how they affect the species in their environment. To do this, we will use species distribution models adapting them, that is, incorporating this exceptionality to the model so that the model is able to absorb this information without detecting these observations as anomalous, but as the extreme events that they are.

In this initial phase of the project, different approaches such as Hurdle models, triangular distributions or others present in the literature have been tested in order to observe advantages and disadvantages of each of them and to be able to provide improvements in the inferential process as well as which covariates of the environment can help to predict or anticipate the harmful effect that an extreme event can cause on certain species populations, such as storms in fish farms or heat waves on birds in wetlands, etc.

All models and their possible improvements will be approached from a Bayesian perspective either by making the study only temporal or by including spatial information with spatio-temporal studies.

Keywords: extreme events, Bayesian modelling, species distribution models.

A hierarchical spatial model for small area estimation of survey-based ordinal variables

Miguel Ángel Beltrán Sánchez[1], *Miguel Ángel Martínez Beneito*[2], *Ana Corberán Vallet*[3]

[1]angel.beltran@uv.es, [2]miguel.a.martinez@uv.es, [3]ana.corberan@uv.es
Departamento de Estadística e Investigación Operativa, Universitat de València

Geographical studies in small areas are an excellent epidemiological tool. Most studies aim to monitor health problems from specific events, such as death counts or disease incidence. Usually, these studies are based on the analysis of information from disease registries or health databases. However, the use of alternative data sources, such as Health Surveys which are periodically collected, allows exploring other health indicators such as mental health, limitations, social support, health habits... These features are usually coded as ordinal variables and their analysis is an important topic in Public Health. Nevertheless, the complex sampling design of many surveys, specifically Health Surveys, makes it impossible to directly apply commonly used models in disease mapping. Hence, it is fundamental to adapt these models for the analysis of survey data, which are usually ordinal.

The methodology proposed here is based on Bayesian hierarchical models, where a categorical likelihood is used at the first level of the hierarchy to describe ordinal data. We apply these models to the analysis of the Health Survey of the Valencian Community in 2016 (HSVC2016) to describe the geographical distribution of different health indicators of interest in this region. Specifically, this work presents and interprets the maps for the main health habits. Through the proposal and mapping of synthetic measures for each question of the survey, the data can be easily summarized and exploited to a greater extent. These results can be used by health agencies to make better decisions or intervene specifically in those areas of the region with lower health levels.

Keywords: Disease mapping; survey analysis; ordinal data analysis.

A new score test for distinguishing between the zero-inflated Poisson and the two-component Poisson mixture distribution

Anabel Blasco-Moreno[1,2], *Pere Puig*[2]

[1]anabel.blasco@uab.cat, Servei d'Estadística Aplicada, Universitat Autònoma de Barcelona
[2]ppuig@mat.uab.cat, Department of Mathematics, Universitat Autònoma de Barcelona

Many disciplines, including medical, biology, genetics, economics, and social sciences, require an understanding of the underlying nature of the data. Data such as the number of days spent in the hospital, the number of chromosomal aberrations, or the number of standard alcoholic drinks are non-negative and often right-skewed, heavy-tailed, and multi-modal, with a point mass at zero. The presence of these features can make it difficult to choose the best distribution for fitting the data.

In biological dosimetry, for example, exposure to ionising radiation (IR) causes a variety of damage in peripheral blood limphocytes. The microscopic counting of dicentric chromosomal aberrations is used to estimate the individual absorbed radiation dose. We look at two types of IR that are important for medical diagnosis and treatment: whole body irradiation (WBI) and partial body irradiation (PBI). Furthermore, radiation exposure might be homogenous or heterogeneous. The number of dicentric chromosomes per cell in a homogeneous PBI scenario can be explained by a Zero-Inflated Poisson distribution (ZIP). In WBI, heterogeneous irradiation scenarios can be modeled by a finite mixture of Poisson distributions (MP). Knowing whether the dicentric distribution corresponds to ZIP or MP allows us to determine whether the exposure was homogeneous PBI or heterogeneous WBI and act accordingly.

In this research, we propose an exact test to contrast the null hypothesis H_0 : Data are ZIP distributed, against the alternative hypothesis H_1 : Data follow a two-component Poisson mixture distribution. Our score test was developed as a solution to a problem involving occupancy distributions. We work with the conditional probability of the data given sufficient statistics, i.e., statistics that contain all of the information about the ZIP distribution's parameters, that is, the whole sum of the observed values and the number of zero-counts. Given a sample of n independent observations of counts $\mathbf{X} = (X_1, X_2, ...X_n)$ following a ZIP distribution, where $\sum_{i=1}^{n} X_i$ and N_0 are sufficient statistics, the probability function of $\mathbf{X}$ conditioned to these sufficient statistics is given by the following expression:

$$Pr\left(\mathbf{X} \mid N_0 = n_0, \sum_{i=1}^{n} X_i = t\right) = \frac{n_0! t!}{n! S(t, n-n_0) \prod_{i=1}^{n} X_i!}\,. \tag{1}$$

This result is related to occupancy problems, and $S(t, n-n_0)$ denotes the second order Stirling number. With the values of the sufficient statistics, expression (1) allows for the generation of ZIP data. The null hypothesis can then be tested by determining the empirical distribution of the score test statistic, which is independent of the parameters. Because the obtained distribution of the test statistic is not asymptotic, it is an exact test. Finally, we applied the score test to several application examples based on in vitro data of heterogeneous WBI and homogeneous PBI.

Keywords: zero-inflated Poisson, mixture-Poisson, score test.

Bayesian variable selection with missing data: An application to cardiology

Stefano Cabras[1], *María Eugenia Castellanos*[2], *Anabel Forte*[3], *Gonzalo García-Donato*[4], *Alicia Quirós*[5]

[1]stefano.cabras@uc3m.es, Department of Statistics, Universidad Carlos III de Madrid
[2]maria.castellanos@urjc.es, Department of Informatics and Statistics, Universidad Rey Juan Carlos
[3]anabel.forte@uv.es, Department of Statistics and OR, Universitat de Valencia
[4]gonzalo.garciadonato@uclm.es, Department of Economy and Finance, Universidad de Castilla-La Mancha
[5]alicia.quiros@unileon.es, Department of Mathematics, Universidad de León

Linear regression models are widely used in medical studies to identify the factors that influence a particular outcome or medical condition. Most commonly, data will have some missing values, and rarely is the missing mechanism completely at random. In medicine, the main methods of dealing with missing data is to eliminate cases with incomplete data or impute it using the mean of each variable. However, these may lead to inconsistency in the analyses and the potential alteration of the relationships between the response and the explanatory variables.

The main challenge faced by many researchers is variable selection with missing data, a topic that has received very little attention in the literature. Current proposals use multiple imputation to complete the database and Rubin's rules to combine the results, but their theoretical underpinnings are still unclear.

To investigate this, we propose an application of a new Bayesian variable selection method for linear regression models in the presence of missing data. We analysed the dataset of the PREDICT-MVI study, which aimed to test whether certain physiological indices calculated during the intervention for acute myocardial infarction could predict the extent of microvascular damage caused by the infarction. The dataset had a 63% of missing data, meaning that there were 22 complete observations, out of the 60 patients recruited.

The employed method reliably calculates Bayes factors (and thus the model posterior probabilities) for all possible models by performing a Monte Carlo approximation of the data densities in each model (known as marginals), which takes into account the variability due to data imputation. The benefits of our approach are: i) A Bayesian method with a probabilistic foundation; ii) The results of variable selection are directly interpretable as they are the probabilities that those are included into the true model; iii) All collected data are used and no subjects are discarded; iv) The method is readily available as it uses already implemented R packages.

Keywords: Model posterior probabilities; Objective priors; Myocardial Physiology indices.

Development of imaging biomarkers for ALS (Amyotrophic Lateral Sclerosis) using multivariate statistical techniques and machine learning

Carot-Sierra, J.M.[1], *Gil-Chong, P.*[2], *Vázquez-Barrachina, E.*[3], Cerdá-Alberich, L.[4]

[1]jcarot@eio.upv.es, Departamento de Estadística e Investigación Aplicadas y Calidad, Universitat Politècnica de València

[2]pgilchong@gmail.com, Departamento de Estadística e Investigación Aplicadas y Calidad, Universitat Politècnica de València

[3]evazquez@eio.upv.es, Departamento de Estadística e Investigación Aplicadas y Calidad, Universitat Politècnica de València

[4]leonor_cerda@iislafe.es, Grupo de Investigación Biomédica en Imagen, Instituto de Investigación Sanitaria La Fe

Amyotrophic Lateral Sclerosis is a degenerative motor neuron disease characterized by its diagnostic difficulty: more than 90% of cases are sporadic and there is no reliable paraclinical test capable of detecting it. The development of ALS biomarkers for diagnosis and monitoring is urgently needed.

This work has used a dataset of 211 patients (114 ALS, 45 mimic, 30 genetic carriers and 22 control) with radiomics attributes (morphometry, iron deposition) integrated with clinical variables and 6 semiquantitative visually-assessed indicators of iron deposition.

A binary classification task approach has been taken to classify patients with and without ALS. A sequential modelling methodology, understood from an iterative improvement perspective, has been followed. It has included variable filtering techniques, dimensionality reduction techniques (PCA, kernel PCA), oversampling techniques (SMOTE, ADASYN) and classification techniques (logistic regression, LASSO, Ridge, ElasticNet, Support Vector Classifier, K-neighbours, random forest). For each proposed architecture, several subsets of the available data have been used, proposing models with single datatypes and multimodal models.

The best results have been provided by a voting classifier composed of five classifiers: accuracy=0.896, AUC=0.929, sensitivity=0.886, specificity=0.929. The best results without the use of semiquantitative variables have been provided by Support Vector Classifier: accuracy=0.815, AUC=0.879, sensitivity=0.833, specificity=0.794. In both classifiers a filtering of variables by feature importance in LASSO has been used.

Keywords: biomarker, radiomics, iterative modelling

Fixed and Random Effects Selection in Generalized Linear Mixed Effects Models

Danae Carreras-Garcia[1], *Ana Arribas-Gil*[2], *David Delgado-Gomez*[3]

[1]dcarrera@est-econ.uc3m.es, [2]aarribas@est-econ.uc3m.es,[3]ddelgado@est-econ.uc3m.es,
Department of Statistics, University Carlos III of Madrid

Generalized linear mixed effects models (GLMM) are widely used in the analysis of correlated or clustered data, such as longitudinal data for repeated measurements. They allow the inclusion of subject-specific parameters via random effects and population characteristics through fixed effects. In addition, GLMM are able to model non-continuous outcomes, which extends both generalized linear models and linear mixed models.

Maximum Likelihood estimation in GLMM has been widely discussed because the likelihood function involves an N-dimensional integral which usually can not be integrated out explicitly. A second issue lies in the fact that the covariance matrix must be positive definite, which leads to ill conditioned problems in estimation procedures when the matrix is close to singular. Another important question in GLMM is variable selection. That is the choice of a model that minimizes a certain criterion based on a trade-off between model fit, and model complexity. Most variable selection procedures are based on the inclusion of a penalization for the parameters in the optimization function. However, the majority of these methods only select significant fixed effects because in order to remove a random effect, an entire row and column of the matrix must be removed.

The goal of this study is to simultaneously select relevant fixed and random effects in GLMM through penalization. We adopt a Cholesky decomposition of the covariance matrix, that ensures the positive definiteness, thus leaving the estimation problem unconstrained. We approximate the N-dimensional integral using the Laplace approximation, and optimize the parameters with the Iterative Soft Thresholding Algorithm with Stochastic Coordinate Descent actualizations.

Keywords: Variable Selection, Regularization, Generalized Linear Mixed Effects Models.

Evaluation of management plans for almond leaf scorch disease in Alicante

Martina Cendoya[1], *Elena Lázaro*[1], *Ana Navarro-Quiles*[2], *Antonio López-Quílez*[2], *David Conesa*[2], *Antonio Vicent*[1]

cendoya_marmar@gva.com
[1]Centre de Protecció Vegetal i Biotecnologia, Institut Valencià d'Investigacions Agràries
[2]Departament d'Estadística i Investigació Operativa, Universitat de València

The plant pathogenic bacterium *Xylella fastidiosa* is a priority quarantine pathogen in the EU (Commission Implementing Regulation (EU) 2019/2072). In Alicante, Spain, the presence of *X. fastidiosa* was first reported in 2017, as the cause of almond leaf scorch disease (ALSD). Based on the legislation that establishes specific measures against this quarantine pathogen (Commission Implementing Regulation (EU) 2020/1201), under an outbreak situation a buffer zone around the infested zone is delimited, where intensive surveillance and control measures have to be implemented. From an individual-based epidemiological model, the ALSD spread in the affected area of Alicante was simulated, and different survey designs and control measures were implemented to compare their effectiveness on the outbreak management. In the disease spread model, the infection of susceptible individuals depends on the transmission rate of infected individuals and the spatial dependence between them through the Matérn correlation function. The survey design was based on the European Food Safety Authority (EFSA) guidelines for statistically sound and risk-based surveys of *X. fastidiosa*, where the survey effort, i.e. sample size, was estimated based on the hypergeometric distribution. One-step and two-step approaches to survey design were compared with different confidence levels for both approaches, including those set by the legislation. Different sizes of buffer zone and eradication radius were also evaluated, with a buffer zone of 2.5 and 5 km, and an eradication radius of 50 and 100 m, where the smaller of each is the minimum set by current legislation. In addition, the effect of vector control, including treatments and inoculum reduction, was considered. They were implemented assuming reductions of the transmission rate by 50% and 90% in the buffer zone, and were compared with the baseline scenario of no reduction. It was found that regardless of the survey design, size of the buffer zone and eradication radius, the reduction of the transmission rate had a strong effect in decreasing substantially the number of infected almond trees. Even when the eradication measures were applied, without this reduction of the transmission rate the resulting number of infected trees was similar to those obtained without any intervention. No major differences were observed with the combinations of buffer zone size and eradication radius. Regarding the different survey designs, a higher confidence level resulted in larger survey efforts and a higher efficiency in reducing the number of infected almond trees. Nevertheless, the survey effort had to be very high to remove all infected trees. Although the two-step approach resulted in a higher survey effort compared to the one-step approach with a similar number of hectares inspected, there were no major differences in the results related to the disease spread management.

Keywords: Individual-based model, outbreak management, simulation

Overdispersed Nonlinear Regression Models

Edilberto Cepeda-Cuervo[1], *María Victoria Cifuentes*[2]

[1]ecepedac@unal.edu.co, Department of Statistics, Universidad Nacional de Colombia
[2]mvcifuentesa@unal.edu.co, Department of Statistics, Universidad Nacional de Colombia

In this talk we propose nonlinear regression models in the biparametric family of distributions. In this class of models we propose two new classes of overdispersed nonlinear regression models: the first, defined from the overdispersion family of distributions proposed by Dey, Gelfand and Peng, and the second from a class of compound distributions. For these models, we develop a Bayesian method in which samples of the posterior distributions are obtained by applying an iterated Metropolis-Hastings algorithm obtained by assuming two groups of parameters, defined by the mean and dispersion regression structures. In the first subclass of models, to improve the performance of the iterated Metropolis-Hastings algorithm, we develop worked variables from the Fisher scoring algorithm, through maximum likelihood estimation of the parameters, to build the kernel transition function. A Bayesian method to fit compound models also is proposed. Finally, we present a simulation study and an application to the neonatal mortality to illustrate the use of the proposed models and the performance of the Bayesian method to fit the proposed models.

Keywords: Overdispersed nonlinear regression, Bayesian methods.

An asymptotic lack-of-fit test for multiple quantile regression

Mercedes Conde-Amboage[1], *César Sánchez-Sellero*[2]

[1]mercedes.amboage@usc.es, Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela

[2]cesar.sanchez@usc.es, Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela

Let us suppose that a response random variable Y may depend on several explanatory variables, denoted by a vector X. Then we can write $Y = q_\tau(X) + \varepsilon$, where q_τ is a regression function reflecting the effect of X on Y and ε is an error term. In quantile regression, it is assumed that $P(\varepsilon < 0|X) = \tau$ with $\tau \in (0, 1)$, and then $q_\tau(X)$ is the conditional τ-quantile of Y given X.

In this work we are going to propose a new method to test whether the regression function belongs to a parametric model, that is, to test a null hypothesis $H_0 : q_\tau \in \{q_\tau(\cdot, \theta) : \theta \in \Theta \subset R^q\}$ where θ is an unknown parameter that can be estimated from a sample of (X, Y), say $(X_1, Y_1), \ldots, (X_n, Y_n)$.

The new test is based on the idea that under the null hypothesis the indicators $Z_i = I(Y_i - q_\tau(X_i, \hat{\theta}) \leq 0), i \in \{1, \ldots, n\}$ should not depend on X, where $\hat{\theta}$ is an estimator of θ. Adjusting a logistic regression model of these indicators on the gradient vector $\frac{\partial q_\tau(X_i,\hat{\theta})}{\partial \theta}$, and denoting the resulting fitted indicators by $\hat{Z}_i = \hat{\beta}' \frac{\partial q_\tau(X_i,\hat{\theta})}{\partial \theta}$, the residuals from this logistic fit satisfy $\sum_{i=1}^n \left(Z_i - \hat{Z}_i\right) \frac{\partial q_\tau(X_i,\hat{\theta})}{\partial \theta} = 0$.

The next step will be to consider some basis functions, $u_1(\cdot), u_2(\cdot), \ldots$, like orthonormalized polynomials or cosine functions, that span the space of all functions of X. Then, taking r functions $u_1, \ldots, u_r$, a score statistic can be constructed as $U_r = \sum_{i=1}^n \left(Z_i - \hat{Z}_i\right) (u_1(X_i), \ldots, u_r(X_i))'$ whose squared norm, normalized by $\tau(1 - \tau)$, gives $S_r = \frac{1}{\tau(1-\tau)} U_r' U_r$. Finally, our test statistic will be $T = \max_{r \in \{1,\ldots,R\}} (S_r - 2r)$, where R is a sufficiently large bound for the number of basic functions. The null hypothesis will be rejected if T takes on a large value.

We have obtained the asymptotic distribution of T under the null hypothesis, which does not depend on unknown parameters, so no bootstrap or resampling techniques are required. Besides the computational efficiency of this asymptotic test, consistency of the test versus alternatives was proven and some simulations were carried out to compare its power with other competing tests. Note also that the proposed test is naturally adapted to continuous and categorical explanatory variables.

The new test is applied to check quantile regression models proposed in the literature to describe the effect of a number of explanatory variables on the infant birth weight. Observe that quantile regression models are more useful than mean regression models to describe the effect on low birth weights.

Keywords: quantile regression, lack-of-fit test.

Antedependence Skew-Normal Linear Models for Longitudinal Data

Corrales-Bossio Martha, *Cepeda-Cuervo Edilberto*[2]
[1]martha.corrales@usa.edu.co, Department of Mathematics, University of Sergio Arboleda
[2]ecepedac@unal.edu.co, Department of Statistic, University Nacional of Colombia

In longitudinal data analysis, the assumption of multivariate normality may be questionable, especially when there are atypical data, when the data exhibit past tails or when there is asymmetric behavior of the data (Lin & Wang 2009). In these cases, the multivariate normal skewed distributions have shown to be efficient in the data analysis (Azzalini & Dalla Valle 1996, Azzalini & Capitanio 1999, Sahu, Dey & Branco 2003).

Thus, considering triangular decomposition of variance - covariance matrix (Macchiavelli & Moser 1997, Cepeda 2001, Cepeda & Gamerman 2005), we propose joint modeling of the localization, scale, autoregressive and skewness parameters, assuming multivariate skew-normal distributions of Azzalini and Sahu.

We present results of the analysis of Chronic Kidney Disease Progression and Transition Probabilities in a Large Preventive Cohort in Colombia to illustrate the use of the proposed models and the performance of the Bayesian method to fit these models. Variability in chronic kidney disease (CKD) progression is a well-known phenomenon that underlines the importance of characterizing the said outcome in specific populations. Our objectives were to evaluate changes in the estimated glomerular filtration rate (eGFR) over time and determine the frequency of dialysis admission and factors associated with this outcome.

Keywords: longitudinal data, skew normal distribution, antedependence models

Empirical power of *CoxCombo* test under uncertain proportional hazards: A simulation study

Jordi Cortés Martínez[1], *Marta Bofill Roig*[2], *Guadalupe Gómez Melis*[3]

[1]jordi.cortes-martinez@upc.edu, Statistics & Operations Research, Universitat Politècnica de Catalunya
[2]marta.bofillroig@meduniwien.ac.at, Center for Medical Data Science, Medical University of Vienna
[3]lupe.gomez@upc.edu, Statistics & Operations Research, Universitat Politècnica de Catalunya

The proportional hazard (PH) premise is assumed in most of the clinical trials with time-to-event endpoints. Under this assumption, the Cox model and the logrank test lead to the most powerful test for comparing survival curves to prove the treatment effect. However, this premise may not always hold true. Alternative approaches, such as the modestly weighted logrank test [1] or the maxCombo test [2] have been proposed to be used in cases where proportionality may not be fulfilled. These approaches are, however, based on weighted logrank tests and lack to provide a clear measure of the effect.

In this work, we propose a new statistical test, *CoxCombo*, which combines several treatment effect measures coming from different Cox models. The goal is to compare the efficacy of an experimental treatment against a control treatment in a trial with a primary survival endpoint. We consider the three Wald statistics Z_C, Z_{WC}, and Z_A, obtained by using a weighted estimation in Cox regression [3], corresponding to testing the effect by means of: simple hazard ratio, average hazard ratio, and average regression effects, respectively. We define then the *CoxCombo* statistic as $Z_{CC} = max\{Z_C, Z_{WC}, Z_A\}$, and calculate its p-value by taking the multivariate distribution of the tests into account. We compare the performance of the proposed test, modestly weighted logrank test and *maxCombo* in terms of the power and type 1 error through a simulation study.

We consider four different scenarios of treatment effect behavior over time: 1) PH; 2) early effect; 3) cross effect; and 4) delayed effect. The last three scenarios imply non-proportional hazards. Piecewise exponential distributions are used to model changes in treatment effect over time. We will use the statistical software R (version 4.1.2) to perform the simulations. Specifically, the `nphRCT`, `coxphw`, and `multicomp` packages will be used to simulate the survival data; to calculate the Wald statistics; and to handle the multiplicity, respectively. In this talk, we discuss the simulation results and the potential of the CoxCombo test as an alternative to the other proposed methods.

1. Magirr D, Burman CF. Modestly weighted logrank tests. Stat Med. 2019;38:3782-90.

2. Lin RS, Lin J, Roychoudhury S, et al. Alternative Analysis Methods for Time to Event Endpoints Under Nonproportional Hazards: A Comparative Analysis, Stat. Biopharm. 2020;12:187-98

3. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratio by weighted Cox regression. Stat Med. 2009;28,2473-2489.

Keywords: Non-proportional hazards, CoxCombo, MaxCombo

Frequentist and Bayesian compositional model to analyse longitudinal microbiome data

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

[1]icreus@alumni.uv.es, Institut de Biologia de Sistemes (I2Sysbio), Universitat de València-CSIC. Departament d'Estadística i Investigació Operativa, Universitat de València.

[2]Andres.Moya@uv.es, Institut de Biologia de Sistemes (I2Sysbio), Universitat de València-CSIC. Fundació per al Foment de la Investigació Sanitaria y Biomèdica de la Comunitat Valenciana (FISABIO). CIBER en Epidemiologia y Salut Pública (CIBEResp)

[3] francisco.santonja@uv.es, Departament d'Estadística i Investigació Operativa, Universitat de València.

Recent studies emphasize the importance of studying longitudinal microbiome data. We have develop a Frequentist and a Bayesian model to analyse the relative abundance of longitudinal microbiome data taken into account the compositional feature of the microbiome data. In both models the relative microbial abundance follow a Dirichlet distribution.

In the Frequentist model, we divide all the Dirichlet parameter between the same selected Dirichlet parameter. The logarithm of this quotient is equal to a regressive expression that use the alr transformation of the taxa and a balance as covariates. The model provides information about the importance of a given taxa and of the rest of the community to define the abundance of that taxa.

In the Bayesian model, the logarithm of the Dirichlet parameters are equal to a regression with the principal balances as covariates. We must mention that principal balances are a compositional tool which chooses the groups of bacteria that maximise the variance and provides information about the relationship between these groups. With this formulation the number of parameters is reduced and we are able to model more microbial taxa. In addition, this model gives information about the dynamics between groups of bacteria are obtained, focusing on which community groups affects the taxa.

We use these two models to analyse microbial taxa time series.

Keywords: Compositional, longitudinal, microbiome.

Semi-parametric generalized estimating equations for repeated measurements in crossover designs

N.A. Cruz[1], *O.O. Melo*[2], *C.A. Martinez*[3]

[1]Corresponding author: neacruzgu@unal.edu.co, Faculty of Nursing, Universidad Nacional de Colombia

[2] oomelom@unal.edu.co, Department of Statistics, Faculty of Sciences, Universidad Nacional de Colombia

[3] cmartinez@agrosavia.co, Corporación Colombiana de Investigación Agropecuaria â AGROSAVIA, Sede Central

Crossover experimental designs with non-Gaussian responses, repeated measures, and complex carry-over effects, i.e., those that depend on both the prior and the affected treatment, are frequent in fields like medicine and animal sciences. Complex carry-over effects feature mathematical intractability under the usual parametric methodologies. Therefore, we proposed a semiparametric model for the analysis of crossover designs with repeated measures within each period of treatment application that accounts for complex carry-over effects. The model was derived using an extension of generalized estimating equations (GEE) with a non-parametric component to model the temporal and carryover effects and a parametric component for the remaining ones; in addition, it can be easily adapted to the case of simple carry-over effects. We considered the usual form of GEE and several spline functions leading to an estimation procedure that is analogous to weighted least squares. Hence, model diagnostics can be performed by adapting the standard procedures of multiple linear regression. Moreover, we established asymptotic results on the estimators that showed a sound theoretical behavior for large samples that was illustrated in simulation exercises. Finally, we compared the proposed methodology with the usual approach in two real datasets: systolic blood pressure and insulin in rabbits, which revealed the advantages of our methodology.

Keywords: Cross-over design, Generalized estimating equations, Kronecker correlation

Are my counts Poisson?

Jacobo de Uña-Álvarez[1], *María Dolores Jiménez-Gamero*[2]

[1]jacobo@uvigo.gal, CINBIO, Universidade de Vigo
[2]dolores@us.es, Department of Statistics and Operations Research, Universidad de Sevilla

The Poisson model is often used to analyze count data. For instance, sequencing experiments in genetics report a large number of read counts along a DNA or RNA region for a (typically) small number of individuals. Several authors have defended the general validity of the Poisson model for the read counts in the referred sequencing setups. Accordingly, multiple testing and other inferences are often performed from Poisson P-values, these are, tail probabilities based on a Poisson model. Still, it is a matter of fact that counts deviate from Poisson in particular applications. In such a case, Poisson tail probabilities are inaccurate and nominal significance levels may be violated.

In this work a test for the null hypothesis that a large number k of (possibly small) samples follow Poisson distributions with arbitrary rates is introduced. The test is based on a differential equation that involves the probability generating function, and that characterizes the Poisson model. The individual test statistics pertaining to the many samples are aggregated into a single measure for which a null Gaussian distribution as $k \to \infty$ is derived. The test may detect any deviation from Poisson when k is large enough (that is, it is omnibus), even for small sample sizes and a vanishing proportion of non-Poisson populations. This is proved both theoretically and through simulations. The method allows for dependences among the samples, which may happen in genome-wide studies. Illustrative applications to real sequencing experiments are provided.

Work supported by the grant PID2020-118101GB-I00, Ministerio de Ciencia e Innovación (MCIN/ AEI /10.13039/501100011033).

Keywords: Goodness-of-fit, High-dimensional data, Multiple testing.

An extension of the individual causal association for continuous non-normal endpoints in a causal inference framework

Gokce Deliorman[1], *Ariel Alonso Abad*[2], *Maria del Carmen Pardo*[3]

[1]gdeliorm@ucm.es, Department of Statistics and O.R., Complutense University of Madrid, Spain
[2]ariel.alonsoabad@kuleuven.be, I-BioStat, KU Leuven, Leuven Belgium
[3]mcapardo@ucm.es, Department of Statistics and O.R., Complutense University of Madrid, Spain

Surrogate endpoints are used to provide earlier informative results on the effectiveness of drug development and new treatment studies in clinical trials. Replacing the true endpoint with a surrogate endpoint is a well-known strategy due to its advantages in reducing the follow-up time and the cost. The evaluation of surrogate variables is complex, and one of the factors that contribute to this complexity is the fact that the true and the surrogate endpoints can be of different natures. Therefore, for each combination of them, complex models and their corresponding surrogate metrics must be developed. The individual causal association (ICA) proposed by Alonso *et al*. (2015) is known to work properly for the normal causal model. In this work, surrogacy is assessed using ICA when both endpoints are continuous and non-normally distributed, and a new approach is suggested to estimate the distribution of the endpoints using non-parametric techniques. The performance of the new approach is analyzed under different scenarios through a simulation study.

Keywords: Individual causal association, Kernel estimation, Surrogate

Improvement of COVID-19 symptoms: a survival analysis study from a Portuguese cohort

Leandro Duarte[1], *Inês Carvalho*[1], *Carla Moreira*[1,2], *Luís Machado*[1], *Ana Paula Amorim*[1], *Joana Costa*[2], *Paula Meireles*[2]

[1]pg45191@alunos.uminho.pt, Centro de Matemática da Universidade do Minho, Universidade do Minho, 4800-058 Guimarães, Portugal

[2]EPIUnit - Instituto de Saúde Pública, Universidade do Porto, Rua das Taipas 135, 4050-600 Porto, Portugal

The COVID-19 pandemic has had a profound impact on the world, affecting millions of people and causing widespread illness and death. As the disease continues, it is critical to understand the patterns and predictors of the disease in order to get valuable information that can be used to develop strategies for preventing and managing COVID-19. Survival analysis techniques have been widely used in medical research to analyze longitudinal time-to-event data, such as time from diagnosis to recovery or death. These techniques provide valuable insights into the risk factors and the outcome of the disease. A registry of 3481 COVID-19 patients diagnosed at Centro Hospitalar Universitário de São João (CHUSJ) between March 01, 2020 and January 01, 2021. Symptoms of the disease were reported at admission, and its improvement was investigated using phone interviews. Descriptive statistics were performed according to the measurement level of the variable, and some nonparametric localization tests were used to compare groups. For the longitudinal analysis, the product-limit estimator of survival (Kaplan and Meier) was used to describe COVID-19-associated symptom duration. The estimated survival curves were used to compare the improvement of COVID-19 symptoms for categorical predictors, and formal hypothesis tests were used. Simple and multiple regression models were used to estimate the effect of potential predictors on the improvement of COVID-19 symptoms.

Keywords: COVID-19, Regression, Survival Analysis.

Statistical models for the analysis of temporal patterns in work-related traffic injuries

Iñigo Elviro[1], *Jesús Asín*[2], *Jorge Castillo-Mateo*[2], *Juan J. Aguilar*[1,3]

[1]720124@unizar.es, jaguilar@unizar.es, Dep. Design and Manufacturing Engineering, University of Zaragoza

[2]jasin@unizar.es, jorgecm@unizar.es, Dep. Statistical Methods, University of Zaragoza

[3] Instituto de Investigación en Ingeniería de Aragón (I3A), University of Zaragoza

Work-related traffic accidents (WTAs) are one of the main causes of injury. WTAs can be divided into two classes, 'in itinere' for commuting and 'in mission' during work time.

In Aragon (Spain), WTAs are the principal cause of sick leave and work-related injuries, with a higher incidence in the industrial parks of the region. Although there are reports that partially summarize the situation, there is no systematic methodology to monitor and analyse patterns in these accidents. This work is part of a collaboration with the Occupational Safety and Health Service of the Government of Aragón (ISSLA) to promote research on this topic.

The aim of the study is to develop statistical models that can be useful to detect and estimate temporal patterns in the events linked to seasonality or light conditions, and other possible causes, such as atmospheric conditions. The response variable is defined as the number of accidents per hour and day in an area. Poisson regression models have been fitted to estimate the WTA rate, considering covariates such as hourly and daily harmonics, location, or meteorological information.

The methodology is applied to analyse the WTA database for the period from 2009 to 2021 in the province of Zaragoza (Aragón). Incremental effects are identified and estimated. Finally, the relative risks are estimated by comparing the rates of the industrial park. The conclusions suggest situations that warn about the need for the authorities to take additional measures to reduce these rates.

Keywords: Commuting risk factors, Shift work related risk, Poisson model.

Bootstrap aggregation for modelling biomarkers' change across the preclinical stage of Alzheimer's disease

Armand G. Escalante[1,2,3], *Marta Milà-Alomà*[1,2,4], *Mahnaz Shekari*[1,2,3,4], *Gemma Salvadó*[1,2,4], *Paula Ortiz-Romero*[1,2], *Juan Domingo Gispert*[1,2,4,5], *Marc Suárez-Calvet*[1,2,4,6], *Natalia Vilor-Tejedor*[1,2,3,7,8]

[1]agonzalez@barcelonabeta.org, Fluid Biomarkers and Translational Neurology/ Neurobiogenetics Team, BarcelonaBeta Brain Research

[2]IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain

[3]Universitat Pompeu Fabra, Barcelona, Spain

[4]CIBERFES, Madrid, Spain

[5]CIBER-BNN, Madrid, Spain

[6]Servei de Neurologia, Hospital del Mar, Barcelona, Spain

[7]Centre for Genomic Regulation (CRG), Barcelona, Spain

[8]Erasmus University Medical Center, Rotterdam, Netherlands.

Background

Whether plasma biomarkers steadily increase during the preclinical stage of Alzheimer's disease (AD) is unknown. Herein, we aimed to determine the rate-of-change of plasma biomarkers throughout preclinical AD. This may be important to determine the optimal time window for treatment.

Methods

We included baseline and follow-up plasma biomarkers measurements (follow-up: 3.37±0.40 years) of 240 cognitively unimpaired participants of the ALFA+ cohort (mean age: 60.75±4.85 years). Plasma Aβ40, Aβ42, GFAP and NfL were measured with the Simoa N4PE Advantage Kit, and plasma p-tau231 with a Simoa-validated in-house assay.For each participant we calculated the difference between follow-up and baseline levels, and corrected for sex, time between measurements, and age. We computed z-scores from the corrected values, and we applied a bootstrapped regression approach to model the rate-of-change of each plasma biomarker as a function of baseline age or Aβ PET centiloids (CL).

Moreover, we also computed the plasma biomarkers rate-of-change in three stages of preclinical AD: (I) CSF/PET Aβ-negative group, (II) CSF Aβ-positive/PET Aβ-negative (CL < 30), and (III) CSF/PET Aβ-positive. Significance was determined if the 95% confidence interval of the rate-of-change did not overlap with zero.

Results

Significant acceleration in the rate-of-change of plasma GFAP was observed with ageing, becoming significant at 60 years. There was also a significant accelerated change of p-tau231 from 55 to 68 years. Plasma NfL was the only biomarker whose rate-of-change accelerated with Aβ accumulation, and that rate became significant at 40 CL. Consistently, we observed a significant acceleration of plasma NfL when there was overt Aβ pathology (CSF/PET Aβ-positive group). The rest of plasma biomarker rates-of-change did not significantly increase, some were even deaccelerating (plasma Aβ42/40 and p-tau231), as Aβ accumulated.

Conclusion

Plasma biomarkers rate-of-change throughout the preclinical AD continuum differ. Plasma NfL rate-of-change accelerates in the later stage of preclinical AD, when overt Aβ pathology is present. The deceleration on the rate-of-change of plasma Aβ42/40 and p-tau231 may explain why their early increase in the preclinical AD continuum tends to plateau in later stages.

Studying the effect of COVID in Suicide-Related Emergency Calls

Pablo Escobar[1], *Miriam Marco*[2], *Antonio López*[3], *María Montagud*[4], *Marisol Lila*[5], *Enrique Gracia*[6]

[1]Pablo.Escobar@uv.es, Department of Social Psychology, University of Valencia
[2]Miriam.Marco2@uv.es, Department of Social Psychology, University of Valencia
[3]Antonio.Lopez@uv.es, Department of Statistics and Operational Research, University of Valencia
[4]Maria.Montagud@uv.es, Department of Social Psychology, University of Valencia
[5]Marisol.Lila@uv.es, Department of Social Psychology, University of Valencia
[6]Enrique.Gracia@uv.es, Department of Social Psychology, University of Valencia

Suicide is a major social and public health problem, accounting for more than 700.000 deaths worldwide every year, according to the World Health Organization (2021). In Spain, suicide has become the leading cause of external death, with 11 people who died by suicide each day (INE, 2022). Several studies have been published studying risk factors both at individual (family history, alcohol abuse, mental illness...) and community level (i.e., social deprivation and social fragmentation). In addition, official data is showing an increasing trend in the last decades and there is a rising concern about the impact of COVID and the restrictive policy measures adopted on mental health. However, data is usually scarce and takes several years to be able to detect the real effects on the population.

Taking all of this into consideration we present a research project focused on studying the effect of COVID in mental health, using suicide-related calls to the 112-emergency service, which depends on the Valencian Regional Government, from 2018 to 2023. This project, which is currently in progress, is focused on using a Bayesian approach, through the Integrated Nested Laplace Approximation (INLA), to understand the spatio-temporal distribution of suicide calls. We will perform different type of analysis:

- **Temporal Evolution:** study of seasonality and trend, group by days, weeks, age and sex.
- **Spatial Distribution:** several spatial models will be implemented, taking into account spatial confounding and age standardization, and studying the effect of community-level risk factors.
- **Spatio-Yemporal Analysis:** spatio-temporal models, with different types of spatio-temporal interactions and mandatory restrictrions will be implemented.
- **Multivariate Spatial Distribution:** taking advantage of the information given by the original database, we will split the data in 2x2x2 categories, according to the type of caller (victim vs witness callers), the gender of the victim (male vs female) and period (pre-COVID vs COVID) to analyze the shared spatial component.

In this session we will present the project, whose preliminary results showing a drastic increase in the number of calls coinciding with the beginning of lockdowns in Spain, which continued to increase after returning to the new normality. There are, however, clear differences between males and females. The future lines of research will also be presented.

Keywords: Suicide, INLA, Disease Mapping

Competitive risk models in early warning systems for in-hospital deterioration: the role of missing data imputation

Juan Carlos Espinosa Moreno[1], *Fernando García García*[1], *Dae-Jin Lee*[2], *María J. Legarreta Olabarrieta*[3], *Susana García Gutiérrez*[3], *Naia Mas Bilbao*[4]

[1]{jcespinosa,fegarcia}@bcamath.org, Basque Center for Applied Mathematics (BCAM)
[2]daejin.lee@ie.edu, School of Science & Technology, IE University, Madrid, Spain
[3]{mariajose.legarretaolabarrieta,susana.garciagutierrez}@osakidetza.eus, Galdakao-Usansolo University Hospital, Research Unit
[4]naia.masbilbao@osakidetza.eus, Galdakao-Usansolo University Hospital, Critical Care Unit

Early Warning Systems (EWS) are useful and very important tools for evaluating the health deteriorating of hospitalised patients, using vital signs (such as heart rate, temperature, etc.) as the main input, based on electronic health records (EHR) which most of the time result in sparse data sets with high rates of missing data. In this work, we aim to study the effect of different imputation techniques on time-to-event (survival) models.

For each case we have patient's sex and age, as well as longitudinal data along the hospitalisation for 7 vital signs (temperature, systolic and diastolic pressure, heart and respiratory rates, oxygen saturation and neurological state). We summarise these longitudinal data with the following central tendency, order and dispersion statistics: maximum, minimum, first observation, last observation, mean, standard deviation, average variance percentage and average derivative, transforming the original variables into a cross-sectional higher dimensional space, that still having missing data problems. Each hospitalisation has two possible final states: clinical deterioration or favourable discharge. Here, we model the time-to-event with competitive risk models taking into account the covariates.

In the Galdakao-Usansolo University Hospital (Basque Country, Spain), a total of 19.602 hospitalisations (lengths of stay at least 24 hours) were collected during the year 2019, of which 852 (4.35%) resulted in deterioration. These data correspond to 55.8% of males and 44.2% of females. We are using a set of imputation methods, such as central tendency statistics (mean and mode), Multiple Imputation by Chained Equations (MICE), Non-Linear Principal Components Analysis (NLPCA) and Random Forest. We evaluate the performances of the imputation methods described before, via root mean square error and conclude the pros and cons of using each one in medical practice. Then, we use Fine and Gray's competitive risk models and the cause-specific Cox proportional hazard regression to model the time-to-event as a function of imputed summarised data. Finally, we evaluate these models employing the traditional and time-dependent area under the ROC curve, for horizon times of 24, 48, 72, 96 and 120 hospitalisation hours.

Keywords: Competing Risk models, Survival models, Data Imputation

A Shiny App for spatial species distribution modeling

Mario Figueira, David Conesa, Antonio López-Quílez

Abstract

In ecology, Species Distribution Models (SDMs) are a statistical tool that has seen a substantial expansion in its implementation over the last two decades. Along with their widespread use, the complexity of the data analysed and the structures of the models used have increased.

This has led to the development of various tools to facilitate the incorporation and use of these new data and statistical methodologies, mostly embodied in new R packages and shiny applications that allow different types of SDMs to be solved. However, the Integrated Nested Laplace Approximation (INLA) approach, which has been increasingly implemented in the field of ecological sciences, has not yet been integrated into an application that can synthesise the complexity of its code into a user-friendly interface for continuous spatial modelling.

To overcome this shortcoming, we present in this work a novel application that allows the use of INLA for those who are not very experienced, or for those with experience who prefer a tool that allows them to carry out an initial analysis quickly, avoiding the process of writing code.

The app allows both geostatistical and preferential modelling. It integrates the complex and hard coding SPDE-FEM (that stands for Stochastic Partial Differential Equation, along with the Finite Elements Method) approach to perform continuous spatial analysis with a visual interface. Moreover, it allows the use of default settings that automate the process or the customisation of a large number of elements that drive the modelling process. In this way, quick initial evaluations or more rigorous studies of the data provided by the user can be carried out, depending on the user's skill and understanding of the fundamentals underpinning the application.

Keywords: INLA, Geostatistics, Preferential models.

Bayesian additive regression trees (BART) applied to global scale species distribution models (SDMs): comparing present and future projections.

Alba Fuster-Alonso[1], *M. Grazia Pennino*[2], *Xavier Barber*[3], *J. Maria Bellido*[2], *David Conesa*[4], *Antonio López-Quilez*[4], *Jeroen Steenbeek*[5], *J. Carlos Baez-Barrionuevo*[2], *Villy Christensen*[5,6] *and Marta Coll*[1,5]

[1]afuster@icm.csic.es, Instituto de Ciencias del Mar (ICM-CSIC),

[2]Instituto Español de Oceanografía (IEO-CSIC),

[3]Centro de Investigación Operativa, Universidad Miguel Hernández (UMH),

[4]Departamento de Estadística e Investigación Operativa (VaBar), Universidad de Valencia

[5]Ecopath International Initiative (EII),

[6]Institute of the Oceans and Fisheries, University of British Columbia

Marine Ecosystem Models (MEMs) have been developed to analyse the past and future dynamics of life in the oceans. One of such efforts is EcoOcean, a complex, mechanistic and spatio-temporal explicit MEM of the global oceans based on a trophodynamic core. EcoOcean requires as inputs the species native ranges and suitable habitats, and for key environmental conditions, species' functional responses and time-varying maps delivered by Earth System Models (ESMs). The different sources of uncertainty in these inputs may influence the validity and accuracy of EcoOcean results. For this reason, our study explores the use of global Species Distribution Models (SDMs) to reduce the uncertainty associated with these inputs.

A promising new alternative to traditional SDMs classification tree methods is the Bayesian Additive Regression Trees (BART). BART is a non-parametric Bayesian regression approach based on a sum-of-trees model. Then, in order to model the presences/pseudo-absences data to obtain EcoOcean inputs, the model applied in this work was as follows:

$$Y_i \sim Ber(\pi_i), \qquad i = 1, \dots, n\,,$$

$$\phi^{-1}(\pi_i) = \sum_{j}^{m} g_j(\boldsymbol{X}, T_j, M_j)\,,$$

where, Y_i is our response variable (presence/pseudo-absence of species) in each observation *i* associated with a Bernoulli probability distribution; π_i is the probability of presence linked to the predictor by a link function ϕ^{-1}; then, g_j is the *j-th* tree of the form $g_j(\boldsymbol{X}; T_j, M_j)$, where m is the total number of trees, $\boldsymbol{X}$ is a vector of multiple covariates, T_j represents a binary tree structure consisting of a set of interior decision rules and a set of terminal nodes, and $M_j = \{\mu_{1j}, \dots, \mu_{jb}\}$ denote a set of parameter values.

Finally, to test BART's capability as an SDM on a global scale. We performed a suitability study for two globally distributed functional groups: 1) marine turtles and 2) tunas. Our results show that BART is a powerful approach to predict the potential distribution of target species, as well as their relationship with key environmental variables, on a global scale, and the outputs obtained are potentially useful for informing EcoOcean.

Keywords: BART, global scale, climate change.

Modeling the propagation of an epidemic in a stochastic SVIS model when a re-vaccination of the susceptible population is considered

Maria Gamboa Perez[1], *Maria Jesus Lopez-Herrero*

[1]mgamboa@ucm.es, Statistics and Data Science Department, Faculty of Statistical Studies, Complutense University of Madrid

This presentation is focused on the use of continuous-time Markov chains (CTMC) to model the transmission of contagious diseases that do not confer permanent immunity. Population is not isolated and in consequence, the spread of infections may result from either coming into contact with infected individuals within the community or non-community members. A proportion of the population receives an imperfect vaccine that fails with a certain probability in the sense that, some individuals that have been previously vaccinated to prevent disease could be infected.

We describe the evolution of the infectious process in terms of a bi-dimensional CTMC representing the number of vaccinated and infected individuals during the epidemic.

The number of immunized individuals decreases over time due to the imperfect vaccine and external source of infection hypothesis, which can lead to the loss of herd immunity. To prevent this, we establish an alarm threshold for the number of protected individuals, which we refer to as the warning level. The viability of a re-vaccination program is evaluated in order to arise vaccine coverage to the initial situation. To achieve that objective we explore the size of the susceptible population when the alarm threshold for vaccinated individuals is reached. We also quantify the time until a re-vaccination program can be launched. We provide theoretical and algorithmic results to obtain statistical characteristics for both random variables and also present some numerical results for the spread of a diphtheria outbreak.

The talk is based on the following paper:

1. Gamboa, M.; Lopez-Herrero, M.J. *Measures to assess a warning vaccination level in a stochastic SIV model with imperfect vaccine*. Studies in Applied Mathematics 2022, 148(4), p.1411-1438. https://doi.org/10.1111/sapm.12479

Keywords: stochastic epidemic model, imperfect vaccine, eligible group

Modelling recurrent fragility fracture events

Esther García-Lerma[1], *Cristian Tebé*[1], *Klaus Langohr*[2], *Guadalupe Gómez Melis*[2]

[1]egarcial @idibell.cat, ctebe@idibell.cat, Biostatistics Unit, Bellvitge Biomedical Research Institute (IDIBELL)

[2]klaus.langohr@upc.edu, lupe.gomez@upc.edu, GRBIO: Research Group in Biostatistics and Bioinformatics, Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya-BarcelonaTech

Background: Previous studies have shown that bone density loss increases with advancing age and fragility fractures are one of the consequences. As this type of fracture is caused by small mechanical forces rather than major trauma, the occurrence of recurrent fractures is common. However, future progression is not considered in most studies, which focus only on the initial event.

Objective: This study aims to evaluate different recurrent event models to identify risk factors of fragility fractures.

Methods: Data were obtained from the EPIC study. A cohort of subjects aged 50-85 years between 2007 and 2017, was extracted from the Catalan Information System of Development of Research in Primary Care (SIDIAP). Patients with less than one year of follow-up or less than 65 years of age on 1st January 2012 were excluded. Data from major fractures (clinical, vertebral, hip, humerus and wrist) were analysed, including data up to the third fracture. Prentice-Williams-Peterson stratified proportional hazards models were adjusted for the time between events, considering death as a competing event. Age was used as the time scale; so that we had left truncation in addition to right censoring. Cause-specific Cox models were adjusted for the three gaps time, between data entry and first fracture, between first and second fracture and between second and third fracture. Proportional hazard assumption was verified graphically by means of the Schoenfeld residuals. For the analysis, we used the package cmprsk of R.

Results: The analysis cohort included 804158 subjects, 52.1% were women and mean age was 75.2 years. 104539 (13.0%) had at least one fracture and 148121 (18.4%) died before the end of the study follow-up. The median time from cohort entry to first major fracture was 13 years and 10 months. Among those with a first major fracture, the median time to a second major fracture was 2 years and 8 months and 2 years to a third major fracture. Early results show that regardless of the order of the fractures, women and people with a previous fracture are at higher risk. In contrast, use of corticosteroids and diabetes are important risk factors for the first fracture, but not for subsequent fractures.

Conclusions: Recurrent fractures are studied with Prentince-Williams-Peterson models. Preliminary results suggest a different role for classic clinical risk factors depending on fracture occurrence time.

Keywords: Survival, Recurrent Events, Fragility Fractures.

Minimum metabolic information for the reconstruction of the evolution of metabolisms

Irene García Mosquera[1], *Bessem Chouaia*[2], *Mercè Llabrés*[3], *Marta Simeoni*[4]

[1]irene.garcia@uib.es, Mathematics and Computer Science Department, University of the Balearic Islands and Health Research Institute of the Balearic Islands (IdISBa),
[2]bessem.chouaia@unive.it, Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari Venezia
[3]merce.llabres@uib.es, Mathematics and Computer Science Department, University of the Balearic Islands and Health Research Institute of the Balearic Islands (IdISBa)
[4]simeoni@unive.it, Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari and European Centre for Living Technology,Venice, Italy

The metabolism involves chemical reactions that link one to the other, creating a complex network of reactions. In this work, we analysed the potential of a simplified representation of the metabolism as a graph, where nodes are metabolic pathways, and there is an edge between two nodes if their corresponding pathways share one or more compounds. We call it an Abstract Metabolic Network (AMN). Our goal was to investigate the extent to which AMNs help discern the different taxonomic groups and capture evolutionary steps. We considered the metabolism of 7141 species stored in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database and performed a large-scale comparison of their AMNs using graph kernels. Specifically, we employed the Vertex Histogram, Shortest Path, Weisfeiler-Lehman and Pyramid Match graph kernels. We performed various experiments, first by considering the whole set of selected species, then by considering all the Eukaryotes and, within Eukaryotes, their kingdoms separately, and finally focusing on Prokaryotes. The comparison results were investigated through exploratory data analysis (heatmaps, multi-dimensional scaling and clustering techniques) as well as machine learning techniques (support vector machine) for prediction analysis. Looking at the experiments, we observe that all the results turned out to be biologically and evolutionary meaningful. Moreover, although performing differently, all the considered kernels reported a similar clustering pattern at a higher taxonomic level, thus suggesting that such patterns are clear and robust. This allows us to state that AMNs can reflect key evolutionary processes within the metabolism. However, it is also clear that, in general, they fail to capture fine-grain differences between species due to the need for more information on the reactions within each metabolic pathway. This result gives rise to new research questions that we would like to address for future work: How to use AMNs to highlight the relevant topological similarities and differences among the metabolism of different species? Is it possible to explore the pathways dependencies of two or more symbiotic species?

Keywords: Graph kernels, multivariate data analysis, support vector machine

Estimation of patient flow in hospitals using up-to-date data. Application to bed demand prediction during pandemic waves

Daniel García-Vicuña[1], *Ana López-Cheda*[2], *María Amalia Jácome*[3], *Fermín Mallor*[4]

[1]daniel.garciadevicuna@unavarra.es, Institute of Smart Cities, Public University of Navarre, Campus Arrosadia, Pamplona, 31006, Spain

[2]ana.lopez.cheda@udc.es, Research Group MODES, CITIC, Departamento de Matemáticas, Universidade da Coruña, A Coruña, 15071, Spain

[3]maria.amalia.jacome@udc.es, Research Group MODES, CITIC, Departamento de Matemáticas, Facultade de Ciencias, Universidade da Coruña, A Coruña, 15071, Spain

[4]mallor@unavarra.es, Institute of Smart Cities, Public University of Navarre, Campus Arrosadia, Pamplona, 31006, Spain

Hospital bed demand forecast is a first-order concern for public health action to avoid healthcare systems to be overwhelmed. Predictions are usually performed by estimating the number of patients admitted to hospital and simulating these inpatients pathways. This requires estimating the distribution of lengths of stay in different hospital facilities such as ward and ICU, and the corresponding branching probabilities. In most approaches in the literature, estimations are parametric and rely on not updated published information or historical data. This may lead to unreliable estimates and biased forecasts during new or non-stationary situations. We introduce a flexible adaptive procedure to estimate efficiently these probabilities and lengths of stay using only near-real-time information of inpatients. The main challenge of using up-to-date patient-level information is that data provided by patients still in hospital ward at the time of estimation is censored, as the future path of these patients remains unknown. We show that methods that take advantage of the partial information associated to these patients using mixture cure models (MCM) are more efficient that naive methods that do not use survival analysis techniques. This is very relevant, for example, at the first stages of a pandemic, when there is much uncertainty and too few patients have completely observed pathways. The performance of the proposed method is assessed in an extensive simulation study in which the patient flow in a hospital during a pandemic wave is modelled. We further discuss the advantages and limitations of the method, as well as potential extensions.

Keywords: EM algorithm, length-of-stay estimation, mixture cure model.

References

1. García-Vicuña D., López-Cheda A., Jácome M.A., Mallor F. (2023). Estimation of patient flow in hospitals using up-to-date data. Application to bed demand prediction during pandemic waves. *PLOS ONE* 18(2): e0282331. doi: 10.1371/journal.pone.0282331

Compositional data analysis to explore the association between joint brain volumetric variation and genetic predisposition to Alzheimer's disease

Patricia Genius[1], *M.Luz Calle*[2], *Raffaele Cacciaglia*[1], *Arcadi Navarro*[1], *Juan Domingo Gispert*[1], *Natalia Vilor-Tejedor*[1]

pgenius@barcelonabeta.org, [1]Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation, Barcelona, Spain. malu.calle@uvic.cat, [2]Biosciences Department, Faculty of Sciences, Technology and Engineering, University of Vic-Central University of Catalonia, Vic, Spain. rcacciaglia@barcelonabeta.org,[1]Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation, Barcelona, Spain. anavarro@barcelonabeta.org, [1]Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation, Barcelona, Spain. jdgispert@barcelonabeta.org,[1]Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation, Barcelona, Spain. nvilor@barcelonabeta.org, [1]Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation, Barcelona, Spain.

Genetic characterization of brain volumes (i.e. imaging genetic studies; IG) is often assessed and analyzed using correlation and univariate modelling. However, given the structural heterogeneity of the brain, it is likely that correlated brain regions may vary together in morphometric characteristics, and genetics may play a role in this joint variation. We propose a new modelling strategy for volumetric analysis in IG studies based on the use of compositional data analysis (CoDA).

The CoDA approach focuses on the relative variation between brain features (components) instead of analyzing them separately. In this study, a CoDA methodology (*coda4microbiome* algorithm) was used, which involves modelling and variable selection (elastic net penalized regression) to identify brain structural variation signatures involving brain features that vary together based on the genetic predisposition to specific neurological conditions. Specifically, in this study, the focus was on Alzheimer's disease (AD) and different disease stages.

The study included middle-aged participants from the ALFA (Alzheimer's and FAmilies) and ADNI (Alzheimer's Disease Neuroimaging Initiative) studies who had magnetic resonance imaging scans, genetic information and cerebrospinal fluid-amyloid status. Individuals were classified into different groups based on the disease-stage and amyloid status: cognitively unimpaired amyloid-beta negative (CU) Aβ- (N=220), CU Aβ+ (N=118), (mild cognitive impaired) MCI Aβ+ (N=230) and AD Aβ+ (N=100). We used the *coda4microbiome* algorithm to analyze the joint structural variation of specific brain components (optimal brain signature) that was most closely associated with higher genetic predisposition to AD (Polygenic Risk score of AD; PRS-AD) (High genetic predisposition to AD: PRS-AD $\geqslant$ quantile 0.8). The elastic net penalty term was defined through a cross-validation procedure. After reparameterization, the brain signature was expressed as a weighted sum of the selected variables in the form of a log-contrast function (Equation 1). The joint change of the brain structural variation signature and the genetic predisposition to AD was assessed through disease-stage stratified logistic regression models adjusted for age and sex.

$$\text{(Equation 1) Signature} = \sum_{j=1}^{K} \hat{\theta}_j \cdot \log(x_j), \text{ where } \sum_{j=0}^{K} \hat{\theta}_j = 0$$

Through the application of the *coda4microbiome* algorithm, we found disease-stage specific joint volumetric variations associated with higher genetic predisposition to AD. These variations were related to specific memory network regions of the brain. Our study highlights the potential of implementing CODA methods to address issues in the neuroimaging research area and provides new insights into the genetics of AD.

Keywords: Compositional Data Analysis; Imaging Genetics; Coda4microbiome.

A retrospective analysis of alcohol-related emergency calls to the ambulance service in Galicia

Mª José Ginzo Villamayor[1], *Paula Saavedra Nieves*[2], *Dominic Royé*[3] *and Francisco Caamaño Isorna*[4]

[1]mariajose.ginzo@usc.es, Department of Statistics, Mathematical Analysis and Optimization (USC)
[2]paula.saavedra@usc.es, Department of Statistics, Mathematical Analysis and Optimization (USC) and Galician Centre for Mathematical Research and Technology (CITMAga)
[3]dominic.roye@ficlima.org, Climate Research Foundation (FIC) and Epidemiology and Public Health Networking Biomedical Research Centre (CIBERESP)
[4]francisco.caamano@usc.es, Department of Public Health (USC) and Epidemiology and Public Health Networking Biomedical Research Centre (CIBERESP)

Keywords: Alcohol, bayesian hierarchical models, nonparametric level set estimation.

This work will be focused on the introduction of statistical methods for data processing and modeling in society, specifically, on alcohol consumption and abuse in Galicia. Dataset is available from a retrospective cohort study based on the telephone calls to the Galicia-061 Public Health Emergency Foundation after alcohol consumption from 1 January 2007 to 4 February 2018. Bayesian hierarchical models and nonparametric level set estimation techniques will applied.

The main objective is modeling spatial and spatio-temporal patterns of emergency calls to the department ethyl poisoning in this region. By fixing administrative areas, for example, municipalities, spatial and spatio-temporal methods for counting data can be considered in this setting. This approach allows to allow to study the evolution of callings patterns. Specifically, hierarchical modeling, through Besag York Molliè (BYM) method will be used to meet this goal (see Besag *et al.* (1991) and Rue and Held (2005) for more details). Integrated Nested Laplace Approximation will be considered in order to fit this kind of models. The analysis will be performed by using covariates such as age, gender, study level, Gini index, incomes, number of bars and regulations/sanctions.

Nonparametric level set estimation techniques will be applied in order to identify the hot-spots of emergency calls. Significant covariates detected from hierachical models fittings will be taken into account. In particular, differences between patterns by gender will be studied.

References

Besag J., York J., Molliè A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43(1)**, 1-20.

Rue H., Held L. (2005). Gaussian Markov Random Fields. London: Chapman and Hall, CRC Press.

and Hall.

Inference under a second order Markov model

Guadalupe Gómez Melis[1], *Mireia Besalú Mayol*[2]

[1]lupe.gomez@upc.edu, Dept. d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya-BarcelonaTech (UPC)

[2]mbesalu@ub.edu, Dept. Genètica, Microbiologia i Estadística, Universitat de Barcelona

Multistate modelization (MSM) is a comprehesive approach to study the evolution of individuals along time and through different states. As complementary to standard survival methods where the analysis is centered in the time to one particular event, in MSM a dynamical path across several states is considered. Therefore, disease burden is captured more fully in comparison to conventional survival analysis. MSM models are well developed for continuous and discrete times, under a first order Markov assumption. If the first order Markov condition does not hold, some solutions have been proposed for specific MSM such as the illness-death model with only 3 states.

Motivated by a cohort of more than 3000 COVID-19 patients hospitalized during the first wave of the pandemic (March-April 2020) a complex multistate model for their dynamic evaluation after hospitalization was designed thanks to the fruitful collaboration between biostatisticians and clinicians. Specifically, this MSM is based on 14 possible transitions among the seven following states of a patient: Non Severe Pneumonia (NSP), Severe Pneumonia (SP), Non-Invasive Mechanical Ventilation (NIMV), Invasive Mechanical Ventilation (IMV), Recovery (R), Hospital Discharge (HD) and Death (D). Since a preliminary analysis showed that the first order Markov condition was not met for some transitions, we propose a more general second order Markov model which would lessen some of these restrictions while providing a more realistic description of the reality.

We have developed a second order Markov model where the future evolution not only depends on the current but also on the preceding state. Under a discrete time analysis (days), assuming that past information is restricted to 2 consecutive times and under homogeneity, we have expanded the transition probability matrix to M different matrices of order M (M is the number of states) and have proposed an extension of the Chapman-Kolmogorov equations.

Estimation and inference for the transition probabilities is presented for complete uncensored data. This is used, among others, to compute the transition probability from SP to IMV, after a given number of hospitalized days, and differentiating between patients that arrive to the hospital with SP from those who develop the disease at the hospital. A counting process framework for the multistate model is discussed as a first attempt to define estimators for the transition probabilities under right-censoring.

Keywords: Multistate Models, Non-Markov, COVID-19

Association between Anthropometric Status at Birth and Postnatal Growth Trajectories in Infants: Evidence for Brain Sparing Effect

Tomás González Garello[1], *Gerardo Cueto*[1], *Jimena Barbeito*[2], *Noelia Bonfili*[2], *Paula González*[2], *Pablo Nuñez*[1], *Adriana Pérez*[1]

[1]tomas21.gg@gmail.com, gercueto@gmail.com, Grupo de Bioestadística Aplicada, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

[2] CONICET-Hospital El Cruce Dr. Nestor Kirchner-Universidad Nacional Arturo Jauretche, Unidad Ejecutora de Estudios en Neurociencias y Sistemas Complejos, Buenos Aires, Argentina

Delayed growth at birth has been associated with postnatal developmental plasticity events, where nutrients would be preferentially allocated to maintain brain growth. According to the brain-sparing hypothesis, increasing energy allocation to brain growth would favor an early catch-up of this structure. Previous studies suggest that brain growth is less affected than other tissues in fetuses with intrauterine growth restriction, and maternal nutrient restriction has a greater effect on body mass than on brain size, supporting that brain growth would be relatively more preserved. However, fetal growth impairment has been associated with altered developmental processes leading to structural alterations and poorer cognitive functioning, social achievement, and educational success, suggesting that protection is not always complete and has long-lasting effects on brain development. Postnatal growth studies that adjust to large longitudinal databases to test this hypothesis are scarce. The objective of this study is to evaluate the association between anthropometric status at birth and postnatal physical growth during the first year of life. We expect a faster physical recovery in cases of low head circumference (HC) at birth compared to recovery rate of length in cases of low length at birth. This is a longitudinal design that includes anthropometric records of 3,399 newborns in Argentina, with a minimum of 3 HC and length records during the first year of life. We used the mixed Count model to adjust longitudinal growth data based on age, comparing four categories of birth growth status, separately for each sex. Each category arises from the combination of low/normal levels of HC and length z-score (HCZ and LAZ) at birth. Recovery indicator for HC and length was taken as the time until the mean growth trajectory estimated for each category surpassed the threshold curve predicted by z-score = -2 for age (which is low/normal z-score threshold). There were significant differences between groups in all growth parameters ($p < 0.05$). Within the group with **low** HCZ and low LAZ at birth, the recovery rate of HC was higher than that of length. In the case of low z-score at birth in only one variable, newborns with low HCZ recovered HC more quickly than the recovery rate of length in individuals born with low LAZ. These results suggest that the trade-off between brain and bone growth is relevant in vulnerable infants, especially for those with low HCZ and low LAZ at birth. We found that, during the first postnatal year, those infants who presented signs of intrauterine delayed growth at birth displayed an accelerated recovery of HC growth. This finding is in line with previous studies that found that fetal brain growth is spared under adverse conditions and adds evidence to support that energy and nutrient trade-off are also present postnatally and is determined by the growth status at birth. Future studies assessing maternal nutritional status and diet habits would enrich our understanding of the extent to which specific types of malnutrition impose trade-offs on the growth of the offspring. The study contributes to the understanding of the consequences of environmental conditions on fetal growth and highlights the importance of monitoring fetal growth during critical periods.

Keywords: **Chiild Growth; Catch-up; Count model.**

Network-based R-statistics software for longitudinal designs: application in a fMRI brain scan database

Zeus Gracia-Tabuenca[1], *Sarael Alcauter*[2]

[1]zeus@unizar.es, Department of Statistical Methods, University of Zaragoza
[2]alcauter@inb.unam.mx, Department of Behavioral and Cognitive Neurobiology, National Autonomous University of Mexico

Many biological systems can be modeled as a set of interacting elements, allowing its characterization from a systems perspective using network science. However, this implies a quadratic increases in the variables to take into account and traditional multiple comparison corrections tend to be conservative. To this regard, Zalesky et al. (2010) proposed Network-Based Statistics (NBS), an algorithm that identifies clusters of edges within the network, and tests their family-wise error (FWE) rate generating a null distribution of cluster maximum size via permutation. Nevertheless, NBS relies on the general linear hypothesis testing (GLHT) which limits its application to longitudinal samples, in particular, to unbalanced samples. To overcome this limitation, we developed a publicly available software (https://cran.r-project.org/package=NBR), NBR (Network-Based R-statistics), that performs mixed-effects models (LME) in the NBS framework, allowing the exploration of unbalanced longitudinal samples.

We used NBS and NBR softwares to test GLHT and LME in a publicly available SWU-SLIM database. The dataset includes 333 participants (145 males; 17-28 years old) with two (n=212) or three (n=121) sessions each. All sessions include a resting-state fMRI brain scan and psychometric data. State anxiety scores and brain connectivity network matrices were used. GLHT and LME tested the edgewise brain-behavior relationship for balanced (424 matrices) and unbalanced (787 matrices) samples, respectively. Significance was assessed based on permutation tests including 1000 permutations restricted to within-subject swapping.

The LME approach found a significant subnetwork of brain regions, which includes the cingulum, the frontal, parietal and occipital cortex, and the cerebellum (pFWE = 0.001), while GLHT found no significant results (pFWE = 0.355).

We showed that NBR overpowers GLHT-NBS when dealing with unbalanced longitudinal samples. This is relevant given that missing data is common in longitudinal studies, and balanced testing could dramatically undermine statistical power. Besides, we were able to show a brain network related to anxiety symptoms that vary over time, which would not be identified using standard methods. Considering the growth of longitudinal studies in biological sciences, we anticipate this method being potentially useful in the field.

This research was supported by CONACyT 619683/330142.

Keywords: network based statistics, linear mixed-effects, brain networks

A new methodology for classification of partially observed curves: an application to aneurysm patients

Pavel Hernández-Amaro[1], *Maria Durban*[2], *M. Carmen Aguilera-Morillo*[3]

[1]pahernan@est-econ.uc3m.es, Department of Statistics, University Carlos III de Madrid
[2]mdurban@est-econ.uc3m.es, Department of Statistics, University Carlos III de Madrid
[3]mdagumor@eio.upv.es, Department of Applied Statistics and Operational Research and Quality, Universitat Politècnica de València

Functional data analysis is one of the fastest growing fields in statistical analysis. Modern data sets often consist of complex objects, such as functions. Functional data is usually found as discrete and often noisy observations of the true underlying function, measured at different locations in time, space, or other continuum. In most cases it is assumed that all functions are observed over the full extension of their domain. However, in many real data sets, each curve is observed in a subset of the domain, which may even be different for each curve. This type of data is know as partially observed functional data.

In this work we present a new methodology to fit a generalized scalar-on-function regression model to deal with partially observed functional data. The proposed functional model considers each curve only within its observed subset of the domain; also a penalty is added to the estimation of the functional coefficient in order to control its smoothness trough the smoothing parameter. Additionally a basis representation of the functional data and the functional coefficient of the model is made. This representation allows us to transform our functional model into a mixed effect model and then estimate directly all the model coefficients including the smoothing parameter. We use B-spline basis for our representations but other suitable basis can be chosen.

The performance of the proposed model is tested on a real classification problem from the AneuRisk65 data set (https://statistics.mox.polimi.it/aneurisk/). The goal is to classify each patient into one of two groups depending on the presence and location of the aneurysms, for this classification two functional variables are taken into consideration: the radius and the curvature of the internal carotid artery. Additionally we also test our new methodology in a simulation study and compared its performance with other classification techniques for partially observed functional data.

Keywords: Partially observed functional data, generalized scalar-on-function regression model, B-splines.

Reference standards based on statistical methods for human identification in Mexico: Forensic Science Apps

N. Sofía Huerta-Pacheco[1], *Ivet Gil-Chavarría*[2], *Chantal Loyzance*[2], *Mirsha Quinto-Sánchez*[2]

[1] nshuerta@enacif.unam.mx, CONACYT - National School of Forensic Sciences, National Autonomous University of Mexico

[2] ivetgil@enacif.unam.mx, chantal.loyzance@enacif.unam.mx, mirsha@enacif.unam.mx, National School of Forensic Sciences, National Autonomous University of Mexico

The statistic in forensic science provides the scientific basis for a fundamental area known as 'human identification' in which different disciplines, such as dentistry, anthropology, dactyloscopy, and genetics, to mention a few, participate in the biological profile construction. Nevertheless, this profiling that provides relevant information such as gender, age and ancestry depends directly on the reference standards of the specific population, which so far are not contextualized to the population of Mexico.

Due to this, through different data collections and protocols established in each discipline, qualitative and/or quantitative information is collected, which allows one to give an estimate or approximation to an unknown value that are mainly evaluated by several statistical techniques or methods such as multivariate methods, non-parametric tests, confidence intervals, generalized linear models, among others.

In this work, we present some free distribution applications that have been developed in Shiny that could be contribute not only to research or education but also to forensic practice since new methodological processes and adaptations to protocols based on the current statistical evidence of the Mexican population have been proposed.

Keywords: Shiny apps, multivariate statistics, non-parametric methods, statistical modeling, reference standards, human identification

A study on group bias in healthcare outcomes for nursing home residents during the COVID-19 pandemic in the Basque Country

Hristo Inouzhe[1], *Irantzu Barrio*[1,2], *María Xosé Rodríguez-Álvarez*[3], *Paula Gordaliza*[1,4], *Itxaso Bengoechea*[5], *José María Quintana*[6,7,8]

[1]BCAM - Basque Cneter for Applied Mathematics, Bilbao, 48009, Spain
[2]Universidad del País Vasco (UPV/EHU), Department of Mathematics, Leioa, 48940, Spain
[3]Universidade de Vigo, Department of Statistics and Operations Research, Vigo, 36310, Spain
[4]Universidad Pública de Navarra, Department of Statistics, Informatics and Mathematics, Pamplona, 31006, Spain
[5]Hospital Galdakao-Usansolo, Hospital at Home Unit, Galdakao, 48960, Spain
[6]Hospital Galdakao-Usansolo, Unidad de Investigación,Galdakao, 48960, Spain
[7]Red de Investigación en Servicios Sanitarios y Enfermedades Crónicas (REDISSEC), Galdakao, Spain
[8]Red de Investigación en Cronicidad, Atención Primaria y Promoción de la Salud (RICAPPS), Bizkaia, Spain

We explore the effect of nursing home status in healthcare outcomes such as hospitalisation, mortality and mortality intra-hospital. Amnesty International, among others, claims that in some Autonomous Communities (geopolitical divisions) in Spain elderly people in nursing homes had restrictions in access to hospitals and treatments. Among the general public, this raised an outcry over the fairness of such measures. In this work, the case of the Basque Country is studied under a rigorous statistical approach and a physician's perspective. As fairness/bias is hard to model mathematically and has strong real world implications, this work concentrates on the following simplification: establishing if nursing home status has a direct effect on healthcare outcomes in the presence of other meaningful covariates related to age, patient comorbidity, period of the pandemic, and others. The methods followed here are a combination of established techniques as well as new proposals from the fields of causality and fair learning. The idea behind these methods is as follows: based on the adjustment variables considered, the sample is optimally trimmed so that the groups of residents and non-residents are made as similar as possible. Thus, once the effect of these variables has been mitigated, the impact of nursing home status on healthcare outcomes is studied. The current analysis suggests that as a group, people in nursing homes were significantly less likely to be hospitalised, and considerably more likely to die, even in hospitals, compared to their counterparts during most of the pandemic. Further data collection and analysis is needed to guarantee that this is solely/mainly due to nursing home status.

Keywords: Fairness, Propensity scoring, Causality.

Variable selection with LASSO regression for complex survey data

Amaia Iparragirre[1], *Thomas Lumley*[2], *Irantzu Barrio*[3], *Inmaculada Arostegui*[4]

[1]amaia.iparragirre@ehu.eus, Department of Mathematics, University of the Basque Country (UPV/EHU)

[2]t.lumley@auckland.ac.nz, Department of Statistics, University of Auckland (UoA)

[3]irantzu.barrio@ehu.eus, Department of Mathematics, University of the Basque Country (UPV/EHU) & Basque Center for Applied Mathematics (BCAM)

[4]inmaculada.arostegui@ehu.eus, Department of Mathematics, University of the Basque Country (UPV/EHU) & Basque Center for Applied Mathematics (BCAM)

Complex survey data are becoming increasingly relevant in a number of fields, including social and health sciences. In this framework, the finite population of interest for the study is usually sampled following a complex sampling design, which may include techniques such as stratification, clustering, or a combination of them in different stages of the sampling scheme. In this context, a sampling weight is assigned to each sampled unit, indicating the number of units that this observation represents in the finite population. Due to these particularities, complex survey data do not satisfy independence and identically distributed conditions, and hence, validity of traditional statistical techniques should be checked before applying them to data collected from complex surveys.

LASSO regression models are one of the most commonly used methods for variable selection. In this context, a tuning parameter must be previously selected to fit the models. Cross-validation is the most widely used validation technique in practice to select the optimal value of this parameter in order to minimize the error of the model to be fitted.

Nevertheless, applying LASSO regression models to complex survey data could be challenging for several reasons, including the fact that traditional validation techniques need to be updated in order to work properly with this type of data. In complex survey framework, other approaches, different to the traditional validation techniques are usually used to define partially independent subsets of the sample. Those approaches are known as "replicate weights" methods. However, to our knowledge, they have never been used in a LASSO regression context. The goal of this work is two-fold. On the one hand, we analyze the performance of replicate weights methods to select the tuning parameter for fitting LASSO regression models to complex survey data. On the other hand, we propose new replicate weights methods for the same purpose. In particular, we propose a new design-based cross-validation method as a combination of the traditional cross-validation and replicate weights. The performance of all these methods has been analyzed and compared by means of an extensive simulation study to the traditional cross-validation technique to select the tuning parameter for LASSO regression models. The results suggest a considerable improvement when the new proposal design-based cross-validation is used instead of the traditional cross-validation.

Keywords: complex survey data, LASSO regression, replicate weights.

Multivariate joint analysis of reading habits and practices among the staff of public libraries in Mexico

A. Olivia Jarvio Fernández[1] *and Mario Miguel Ojeda Ramírez*[2]

[1]ojarvio@uv.mx, Centro de Estudios de la Cultura y la Comunicación, Universidad Veracruzana

[2]mojeda@uv.mx, Facultad de Estadística e Informática, Universidad Veracruzana

In this work, a family of "assembled" multivariate techniques is used, which are new joint multivariate analysis procedures, which simultaneously perform two optimization processes: dimensionality reduction (which works on the variables) and grouping (of individuals). from the creation of clusters. The use is illustrated in the framework of a survey sample research that produced data on the reading habits and practices of public librarians in Mexico. The objective of the research was to identify the factors that define a grouping in three segments of librarians to implement training strategies in reading promotion. Two options are implemented: (1) for a series of categorical data, a cluster analysis (CA) is performed together with a dimensionality reduction via multiple correspondence analysis (MCA) ―which here will be called Cluster CA―; and (2) for quantitative data, a cluster analysis (CA) is performed together with a dimensionality reduction by the PCA method, ―which will be called CP Cluster―. For the analysis strategy, the R clustrd library was used, which produces numerical and graphic outputs that report on the association between variables and the grouping of individuals (Markos et al., 2019).

Keywords: Cluster analysis, reading promotion, librarians.

Markos, A., Iodice D'enza, A. and Van Der Velden, M. (2019b): Beyond tandem analysis: Joint dimension reduction and clustering in R. *Journal of Statistical Software (Online***)**, 91(10).

Statistical modeling to adjust for time trends in platform trials utilising non-concurrent controls

Pavla Krotka[1,*], *Martin Posch*[1], *Marta Bofill Roig*[1]

[1]Center for Medical Data Science, Medical University of Vienna
[*]pavla.krotka@meduniwien.ac.at

Platform trials enhance drug development by offering increased flexibility and efficiency. They evaluate the efficacy of multiple treatment arms, with the added benefit of permitting treatment arms to enter the trial over time and to stop early based on interim data. Treatment efficacy is usually assessed using a shared control arm. For arms entering later, the control data is divided into concurrent and non-concurrent controls (NCC), referring to control patients recruited while the given treatment arm is in the platform and before it enters, respectively. Analysis using NCC can reduce the required sample size and increase power, but also lead to bias in the effect estimates and hypotheses tests, if there are time trends.

For platform trials with continuous endpoints without interim analyses, a regression model has been proposed that utilizes NCC and adjusts for time trends by including the factor "period" as a fixed effect. Here, periods are defined as time intervals bounded by any treatment arm entering or leaving the platform. It was shown that this model leads to unbiased effect estimates and asymptotically controls the type I error (T1E) rate regardless of the time trend pattern, if the time trend affects all arms in the trial equally and is additive on the model scale [1]. However, if interim analyses are included, the definition of the factor periods becomes data dependent and the number of periods to adjust for depends on previous interim results. Furthermore, due to early stopping the sample sizes in the different arms become outcome dependent, and therefore treatment effect estimates are no longer unbiased. This can affect the adjustment for time trends in the linear model, and the T1E rate might no longer be controlled.

In this talk, we suggest two extensions of this model. First, we propose an alternative definition of the time covariate by dividing the trial into fixed-length data-independent calendar time intervals. Second, we propose alternative models to adjust for time trends. In particular, we consider: accounting for dependency between closer time intervals by adjusting for autocorrelated random effects; and employing spline regression to model time with a smooth polynomial function. We implement the proposals in the `NCC` R-package [2] and evaluate their performance in terms of the T1E rate and statistical power for individual treatment-control comparisons in a simulation study under a wide range of scenarios.

Keywords: Platform trials, Non-concurrent controls, Statistical modeling, Statistical inference

References

[1] Bofill Roig, M., Krotka, P., et al. (2023). *On model-based time trend adjustments in platform trials with non-concurrent controls.* BMC Med. Res. Methodol.

[2] Krotka, P., et al. (2023) *NCC: An R-package for analysis and simulation of platform trials with non-concurrent controls.* arXiv:2302.12634 (https://arxiv.org/abs/2302.12634).

Smooth $k-$sample tests under left truncation

Adrián Lago[1], *Ingrid Van Keilegom*[2], *Juan Carlos Pardo-Fernández*[3], *Jacobo de Uña-Álvarez*[4]

[1]adrian.lago@uvigo.es, Department of Statistics and Operations Research, Universidade de Vigo
[2]ingrid.vankeilegom@kuleuven.be, Department of Decision Sciences and Information Management, KU Leuven
[3]juancp@uvigo.es, Department of Statistics and Operations Research, Universidade de Vigo
[4]jacobo@uvigo.es, Department of Statistics and Operations Research, Universidade de Vigo

Left truncation arises in many different applied fields due to the impossibility of observation of every individual that experiments the event of interest, frequently as a result of the way a study is designed or limitations on the measurement instruments. Truncation causes an observational bias which also induces bias in the estimators of different population quantities, such as the survival function and, as a consequence, on the estimation of the density function. This implies the necessity to adapt the density function estimator for complete data to left-truncated data. Let us now consider k different populations in which the target variable is left truncated. A common applied problem is to determine whether these target variables follow the same distribution in each of the k populations. To address this problem, a test based on a L_2 distance involving the estimator of the density function in every sample and in the pooled sample is proposed. Its asymptotic null distribution is studied and, due to the difficulty to apply it in practice, a bootstrap resampling plan is proposed to approximate the null distribution of the test statistic. As the test is based on the estimation of the density function, the bandwidth plays an important role on its performance. This leads to propose a choice of the smoothing parameter, based on a double bootstrap algorithm, to maximize the power of the test. The performance of the bootstrap and the choice of the bandwidth will be studied through Monte Carlo simulations. The proposed test will be compared to other test in the literature for left-truncated data, such as the Kolmogorov-Smirnov and the log-rank, under different simulation scenarios to determine under which conditions each one is more adequate to be used. A dataset regarding pregnancy times will be employed to exemplify the performance of those three tests, which will all determine that a drug called coumarin does not have an effect on the pregnancy time until an spontaneous abortion.

Keywords: Left truncation, $k-$sample problem, bootstrap.

Semi-Markov multistate models to analyze the disease progression of hospitalized COVID-19 patients during the first three waves in the Barcelona metropolitan area

Klaus Langohr[1], *Xavier Piulachs*[1], *Natàlia Pallarès*[2], *Carlota Gudiol*[3], *Cristian Tebé*[2], *Guadalupe Gómez Melis*[1]

[1]klaus.langohr@upc.edu, xavier.piulachs@upc.edu, lupe.gomez@upc.edu, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya · BarcelonaTech, Barcelona, Spain
[2]npallares@idibell.cat, ctebe@idibell.cat, Institut d'Investigació Biomèdica de Bellvitge, Barcelona, Spain
[3]cgudiol@bellvitgehospital.cat, Hospital Universitari de Bellvitge, Barcelona, Spain

Data of more than 4000 hospitalized COVID-19 patients from the Barcelona metropolitan area during the first three waves of the corona virus pandemic are the basis of the present work, which is part of the DIVINE project (***D**ynamic evaluat**I**on of COVI**D**-19 cli**N**ical stat**E**s and their prognostic factors to improve the intrahospital patient management*). The work focuses on properly addressing two specific features in the setting of multistate models. First, the traditionally assumed Markov property is subject under examination by using a recently-published Cox-model-based procedure for testing the Markov assumption in a given transition. This test departs from previous approaches in that it provides assessment for virtually any Markov-free multistate modeling context. Second, the multicohort nature of the data set is modelled in two different ways: on the one hand, cohort-specific fixed effects are included in the transition-specific Cox model. On the other hand, stratified Cox models with cohort-specific baseline hazard functions are fitted to the data. This cohort effect, considered either in isolation or interacting with any covariate, is expected to capture underlying disease patterns that are not explicitly collected by other explanatory covariates. The main objective of our proposed models is to provide a general procedure to analyze complex disease processes that may sequentially affect different population cohorts.

The multistate model under study considers two initial states (no severe pneumonia and severe pneumonia), three transient states (recovery; noninvasive mechanical ventilation; and invasive mechanical ventilation), and two absorbing states (discharge and death), and a total of 14 transitions between two subsequent states. The transition-specific hazards are modeled with semi-parametric Cox models that not only account for the effect of transition-specific baseline covariates, but also for the potential impact of the sojourn time at a previous state what converts them in semi-Markov models.

We discuss the strengths and inconveniences of the semi-Markov models and provide graphical tools to illustrate the cohort effect. These added features within our models allow for a better understanding of the biology underlying any pathological process presented in form of sequential cohorts.

Keywords: Stratified Cox model, semi-Markov multistate model.

Bayesian zero-inflated multi-state cure models via INLA

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

[1]llopis_fracar@gva.es, HSRP Unit, Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO), Spain.

[2]carmen.armero@uv.es, Department of Statistics and Operations Research, Universitat de València, Spain.

[3]sanfelix_gab@gva.es, HSRP Unit, Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO), Spain.

Multi-state models are a class of stochastic models in which patients are allowed to move among some states with the time. One particular model, the so-called illness-death model results especially useful to approach many scenarios where death and a main disease or health condition appear together. In its easiest version it involves three states: initial, disease or progression, and death. However, some patients might cure from disease, not being expected to progress or worsen anymore, having a cure model setting. Patients could also die at the beginning of the study, which would lead to zero-inflation. Those epidemiological problems are approachable using a mixture of illness-death models and logistic regression, the first modelling state transitions and the latter modelling a cure or a zero probability. We illustrate the application of zero-inflated multi-state cure models with both a real study involving recurrent hip fracture and death, and with a simulated dataset. Our proposal uses a Bayesian methodology through the integrated nested Laplace approximation (INLA).

Keywords: cure rate, illness-death models, integrated nested Laplace approximation.

Generalised additive model applied to principal component analysis of geographic data

Francisco de Asis Lopez[1],*Javier Roca-Pardinas*[2], *Celestino Ordonez.*[3].

[1]franciscoasis.lopez@uvigo.es,Department of Statistics and Operational Research. University of Vigo
[2]roca@uvigo.es, Department of Statistics and Operational Research. University of Vigo
[3]cgalan.uniovi@gmail.com, Department of Mining Exploitation and Prospecting. University of Oviedo

Abstrat

Geographically Weighted Principal Component Analysis (GWPCA) is an extension of classical PCA to deal with the spatial heterogeneity of geographical data. This heterogeneity results in a variance-covariance matrix that is not stationary but changes with the geographical location. Despite its usefulness, this method presents some unsolved issues, such as finding an appropriate bandwidth (size of the vicinity) as a function of the retained components.

In this work, we address the problem of calculating principal components for geographical data from a new perspective that overcomes this problem. Specifically we propose a scale-location model which uses generalized additive models (GAMs) to calculate means for each variable and a variance-covariance matrix that relates the variables, both depending on the spatial location. This approach does not require to calculate an optimal bandwidth as a function of the number of components retained in the analysis. Instead, the covariance matrix is estimated using smooth functions adapted to the data, so the smoothness can be different for each element of the matrix.

The proposed methodology was tested with simulated data and compared with GWPCA. The result was a better representation of the data structure in the proposed method. Finally, we show the possibilities of our method in a problem with real data regarding air pollution and socioeconomic factors.

Monitoring intimate partner violence against women with spatio-temporal hierarchical structures

Antonio López-Quílez[1], *Pablo Escobar*[2], *María Montagud*[2], *Miriam Marco*[2], *Marisol Lila*[2], *Enrique Gracia*[2]

[1]Antonio.Lopez@uv.es, Department of Statistics and Operations Research, University of Valencia
[2]Pablo.Escobar@uv.es, Maria.Montagud@uv.es, Miriam.Marco2@uv.es, Marisol.Lila@uv.es, Enrique.Gracia@uv.es, Department of Social Psychology, University of Valencia

Intimate partner violence against women (IPVAW) is the most common form of violence experienced by women, with a global prevalence estimated between 22% and 30%. In Spain, according to the latest national survey on IPVAW conducted in 2019, 14.2% of all women aged 16 or older have suffered physical and/or sexual IPVAW at some point in their lives. Despite the wide range of measures that have been taken at the international and national levels to reduce the high prevalence of IPVAW, progress in the reduction of IPVAW is still slow and insufficient. Therefore, new approaches and more advanced tools are needed for more effective responses to reduce the prevalence of IPVAW in our society. We propose the development and implementation of an advanced epidemiological monitoring system for the surveillance and risk prediction of IPVAW in Spain.

The analytical framework underpinning the proposed epidemiological monitoring system for the surveillance and risk prediction of IPVAW is based on the application of Bayesian spatio-temporal models to analyze geographic patterns and temporal trends of risk in small area ecological studies. A software package based on such an analytical framework and using geocoded national-level data on reported IPVAW cases would allow the continuous analysis and visualization of the relative risk of IPVAW and its evolution over time in Spain, taking into account as a minimum unit of analysis all the census block groups in which the country is divided (36,382), as well as their aggregation into broader geographical units (municipalities, regions and autonomous communities). This tool would also make it possible to analyze the relationship between IPVAW risks and the socio-demographic characteristics of the geographical areas analyzed (e.g., urban vs. rural, demographic indicators, socioeconomic indicators) as well as the prediction of future risks at different geographical levels of analysis, and in different time frames (e.g., monthly, quarterly, annual).

A Bayesian spatio-temporal autoregressive model to account for the spatial and temporal dependence of IPVAW risks, allows to deal with issues such as spatial and temporal autocorrelation, overdispersion, or small counts. The impact of various real and simulated social intervention scenarios can be assessed through the application of the monitoring system, by establishing the temporal, geographic and demographic scope affected. Spatio-temporal analysis provides short-term predictions for the next time periods. The usefulness of these predictions is linked to the accuracy of the predictions and the magnitude of their uncertainty, which in turn will depend on the geographical and temporal scale used.

Keywords: Spatio-temporal models, Bayesian inference, IPVAW, Short-term prediction.

Unbiased estimators of kappa coefficients for two raters

Martín Andrés, A.[1] *and Álvarez Hernández, M.*[2]

[1]amartina@ugr.es, Biostatistics, Faculty of Medicine, University of Granada, Spain.
[2]maria.alvarez@cud.uvigo.es, Defense University Center at the Spanish Naval Academy, Spain.

It is often necessary to assess the degree of concordance or agreement between two raters which independently classify n subjects within $K \geq 2$ nominal categories. As some of the observed agreements may be due to chance, it is most common to eliminate the effect of chance by defining a *kappa*-type coefficient. The agreement measure has the expression $\kappa = (I_o - I_e)/(1 - I_e)$ where I_o is the observed index of agreements (sum of the observed proportions of agreements) and I_e is the expected index of agreements (sum of the proportions of agreements that would happen if the two raters acted independently). When the categories are ordinal, or when they are nominal and weights are also assigned to the disagreements obtained, the above definition provides a weighted *kappa* coefficient. In any case, it is common to use the Cohen's *kappa*, Scott's *pi*, Gwet's AC1/2, and Krippendorf's *alpha* coefficients, which are obtained according to the definition adopted for I_e. However, all estimators of previous *kappa* coefficients are biased, since they estimate the product of two population proportions through the product of their sample estimators. The first objective of this study is to correct this bias by proposing unbiased estimators. We also provide the variances of these as a function of the variances of the biased estimators, except in the case of the Gwet estimators. The methodology is easy to apply to any other *kappa* coefficient studied but it may be unnecessary when the sample size n is sufficiently large (e.g. $n \geq 100$). In order to prove this, some simulations and data examples are shown. Finally, we demonstrate that the new unbiased estimator of the Cohen's *kappa* coefficient also coincide with the unbiased estimator of Lin's concordance correlation coefficient if the former are defined assuming quadratic weights.

Keywords: Agreement; *kappa*-type coefficient; Lin's concordance correlation coefficient.

Confidence intervals for the length of the ROC curve based on a smooth estimator

Pablo Martinez-Camblor[1],

[1]Pablo.Martinez-Camblor@Hitchcock.org, Department of Anaesthesiology and Biomedical Data Sciences Department, Geisel School of Medicine at Dartmouth

A good diagnostic test should show different behaviour on both the positive and the negative populations. However, this is not enough for having a good classification system. The binary classification problem is a complex task, which implies to define decision criteria. The knowledge of the level of dissimilarity between the two involved distributions is not enough. We also have to know how to define those decision criteria. The length of the receiver-operating characteristic, ROC, curve has been proposed as an index of the optimal discriminatory capacity of a biomarker. It is related not with the actual but with the optimal classification capacity of the considered diagnostic test. One particularity of this index is that its estimation should be based on parametric or smoothed models. We explore here the behaviour of a kernel density estimator-based approximation for estimating the length of the ROC curve. We prove the asymptotic distribution of the resulting statistic, propose a parametric bootstrap algorithm for confidence intervals construction, discuss the role that the bandwidth parameter plays in the quality of the provided estimations and, via Monte Carlo simulations, study its finite-sample behaviour considering four different criteria for the bandwidth selection. The practical use of the length of the ROC curve is illustrated through two real-world examples.

Keywords: Asymptotic distribution; Binary classification problem; Length of the ROC curve.

GAMLSS models to explore the use of health services in community-dwelling older adults, according to frailty

Maider Mateo-Abad[1], *Kalliopi Vrotsou*[2], *Fracisco Rivas Ruiz*[3], *Itziar Vergara*[4]

[1]maider.mateoabad@biodonostia.org, IIS Biodonostia, Grupo Antencion Primaria & RICAPPS
[2]kalliopi.vrotsou@osakidetza.eus, IIS Biodonostia, Grupo Antencion Primaria & RICAPPS
[3]francisco.rivas.ruiz.sspa@juntadeandalucia.es, Agencia Sanitaria Costa del Sol, Unidad de Investigacion & RICAPPS
[4]itziar.vergaramitxeltorena@osakidetza.eus, IIS Biodonostia, Grupo Antencion Primaria & RICAPPS

Frailty in older adults is a predictor of survival and other health outcomes.It is known that other currently used health indicators, like the burden of the disease, and the number of prescribed drugs are related with a greater use of health resources, but they alone are not sufficient in predicting the provision of health care services for order people. Frailty, measured through a functional performance test, could play a key role as a relevant health indicator. Therefore, the aim of this study was to explore the use of a wide range of health services in community-dwelling older adults, according to their frailty condition. Due to the non-normally distributed nature of these kind of data, generalized additive models for location, scale and shape (GAMLSS) were used seeking to find the best fit distribution to fit the data.

The GAMLSS models for count outcomes were performed for the number of contacts with the general practitioner and primary care nurse; visits to specialists; visits to emergency rooms; and hospital admissions. All models were adjusted for sex, polypharmacy and age-adjusted Charlson Comorbidity Index. The health region was included as a random effect, and the follow-up time as an offset parameter. All fixed effects were considered to model all parameters: μ, the location parameter; σ, the scale parameter; and shape parameters, ν, skewness and τ, kurtosis, respectively. For each type of service use, the distribution that best fitted the data was chosen, based on the generalized Akaike information criterion (GAIC). Logistic regressions were also performed for the dichotomous variables.

The results of the best regression models, showed that frailty was significantly associated with the utilization of every health service considered. The negative binomial distribution type II was the distribution that best fitted most of the count outcomes, except for the general practitioner visits for which the zero-inflated negative binomial distribution, type I, was chosen. The highest IRR was observed for the visits to a primary care nurse $1.5(1.3, 1.6)$, indicating that frail individuals visited these nurses almost more times over a year than robust ones. And they were also more likely to visit an emergency room or be hospitalized (OR $1.3(1.0, 1.6)$ and $1.4(1.1, 1.8)$, respectively).

The GAMLSS models allows for examining not normally distributed complex data, identifying statistically significant factors related with the studied outcome, not only for location parameter.

Funding: Carlos III Health Institute (Numbers: PI14/01003, PI14/01905, PI18/01558)

Keywords: GAMLSS models, use of health services, negative binomial distribution

Classification using a joint model of longitudinal data and binary outcomes based on the SAEM algorithm

Cristian Meza[1], *Maritza Márquez* [2], *Rolando de la Cruz*[3], *Claudio Fuentes* [4]

[1]cristian.meza@uv.cl, INGEMAT-CIMFAV, Universidad de Valparaíso, Chile
[2]maritza.marquez@edu.uai.cl, Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Chile
[3]rolando.delacruz@uai.cl, Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Chile
[4]fuentesc@oregonstate.edu, Department of Statistics, Oregon State University, USA

Despite the pervasiveness of binary outcomes in the vast scientific literature, joint models for longitudinal and binary data are not standard. The idea of jointly modeling these data, which share common parameters or covariates that may link them, is because they provide a practical way of evaluating the association between these two types of data, which, in clinical studies, are usually collected jointly on a series of individuals. In this work, we propose a new estimation method based on a stochastic approximation version of the EM algorithm, the so-called SAEM algorithm, to estimate the parameters of a joint model based on an (non)linear mixed effects model for the longitudinal part and a generalized linear model (GLM) for the primary response of interest. We applied our method to two different sets of data from the monitoring of pregnant women. The first dataset that drove our research proposal corresponds to a follow-up study carried out on 173 young pregnant women during the first trimester of pregnancy, in which the concentration values of the β-hCG hormone were measured. At the time of birth, the women were classified into two groups: a normal group, in which those women who had a normal delivery were considered, and an abnormal group, in which only those women who had any complication that would result in a non-terminal delivery along with the loss of the fetus were considered. According to France et al. (1996), they assured that the hormone concentration values vary from women who have normal pregnancies with terminal deliveries to women who have spontaneous abortions or other types of adverse pregnancy outcomes. This association has made it possible to predict (with some uncertainty) the pregnancy outcomes. Then, for this dataset, we analyze the β-hCG hormone levels measured in early pregnancy that can be used to predict normal versus abnormal pregnancy outcomes. The resulting joint model allows us to carry out a classification process that, from this point of view, improves on previous studies. In the second case, we are interested in the predictions of post-molar gestational trophobastic neoplasia (GTN) studied by Dandis et al. (2020). We want to provide updated predictions of postmolar GTN based on age and repeated measurements of human chorionic gonadotropin (hCG). A model for a binary outcome is required since it is unknown when GTN development began. The joint model considered here uses as predictors in a logistic regression model with the status of GTN as the outcome the random effects from the first model, which models the hCG values via a mixed effects model. For this dataset, the use of the longitudinal data improves the predictive ability of the logistic regression model compared to the reduced model using only the age of the pregnant woman as a baseline covariate.

Keywords: Joint model; Mixed effects models; SAEM algorithm

Generalized spatial conditional overdispersion models: Semiparametric extensions proposals

Mabel Morales-Otero[1], *María Durbán*[2], *Vicente Núñez-Antón*[3]

[1]mmoralesote@unav.es, Institute of Data Science and Artificial Intelligence, University of Navarra
[2]mdurban@est-econ.uc3m.es, Department of Statistics, University Carlos III of Madrid
[3]vicente.nunezanton@ehu.eus, University of the Basque Country UPV/EHU

Generalized spatial conditional overdispersion models represent an excellent choice to fit spatial count data, since they are able to account for overdispersion, capture the possible existing spatial correlation, and they are also flexible enough as to allow for the dispersion to vary according to covariates and/or spatial terms. However, when including covariates in the regression structures for these models, we are assuming that the possible existing relationship between each covariate and the predictor is linear, which may not be necessarily the case, since they could be given by another, maybe non linear pattern. In this sense, smoothing methods should be considered in generalized linear models (GLM), so that the linearity hypothesis can be relaxed. Therefore, in this work we propose a semiparametric extension of the generalized spatial conditional overdispersion models that will allow us to capture such non linear relationships. In particular, for the smoothing of such variables, we have specified P-splines in their mixed model representation. We illustrate their usefulness by fitting them to the study of infant mortality rates and mother's postnatal period screening test in Colombia, where we investigate the possible existence of such non linear relationships. In these applications, we have found evidence of a non linear relationship between the mortality rates and the variable representing the amount of resources provided by the government for academic achievement or education.

Keywords: Spatial models, Semiparametric models, P-splines.

A model to predict ceiling of care in COVID-19 hospitalized patients

Pallarès N[1], Inouzhe H[2], Barrio I[2,3], Fernández D[4], Cortés J[4], Langohr K[4], Videla S[5], Gómez Melis G[4], Tebé C[1]

[1]npallares@idibell.cat, ctebe@idibell.cat, IDIBELL, Barcelona, Catalonia, Spain

[2]hinouzhe@bcamath.org, BCAM - Basque Center for Applied Mathematics

[3]irantzu.barrio@ehu.eus, University of the Basque Country UPV/EHU

[4]jordi.cortes-martinez@upc.edu, daniel.fernandez.martinez@upc.edu, lupe.gomez@upc.edu, klaus.langohr@upc.edu, UPC/BarcelonaTech, Barcelona, Spain

[5]svidela@bellvitgehospital.cat, Bellvitge Universitary Hospital, Barcelona, Spain

Background and objective: Therapeutic ceiling of care is the maximum therapeutic effort to be offered to a subject based on age, comorbidities, and the expected clinical benefit in relation to the availability of resources. According to previous data, COVID-19 subjects with a ceiling of care assigned at hospital admission are mainly older, have more comorbidities, and fewer clinical symptoms at baseline than patients without a ceiling of care. The incidence of death, severe pneumonia, and complications is higher in patients with a ceiling of care. Therefore, analysis of hospitalized subjects with SARS-CoV-2 infection should be stratified by ceiling of care to avoid bias and overestimation of outcomes. The ceiling of care decision is not reported in most published COVID-19 cohorts. Our aim is to develop and validate a model to predict the ceiling of care for hospitalized subjects with COVID-19 using information on the demographic and clinical profile of the patients available at the time of hospital admission.

Methods: The data used to develop the model came from an observational study conducted during four waves of COVID-19 (March 2020-August 2021) in 5 centers in Catalonia. Data were sampled 1000 times by bootstrapping. A logistic regression model with ceiling/no ceiling as outcome was fitted for each sample using backward elimination. Variables retained in more than 95% of the models were candidates for the final model. Alternative variable selection methods such as Lasso, CART, and Boruta were also performed to increase the robustness of the final set of selected variables. Discrimination was assessed by estimating the area under the ROC curve (AUC) and calibration by comparing observed versus expected probabilities of ceiling of care by deciles of predicted risk. Validation was performed in the whole cohort and in subgroups of interest. A cohort of patients diagnosed with COVID-19 in the Basque Country will be used to externally validate the results.

Results: A model including age, COVID-19 wave, chronic kidney disease, dementia, hypertension, heart failure, metastasis, peripheral vascular disease, COPD, and ictus had excellent discrimination (AUC=0.89 [95% CI 0.88; 0.90]) and the observed probabilities agreed well with the predicted probabilities. Patients with relevant comorbidities and by deciles of age also showed excellent figures for calibration and discrimination. External validation of the model in a cohort of patients with COVID-19 is ongoing.

Conclusions: Ceiling of care can be predicted from information on the subject's demographic and clinical profile available at hospital admission. Cohorts without information on ceiling of care can use this model to report outcomes in accordance with it and avoid bias, particularly in overestimating the incidence of outcomes in patients without ceiling of care.

Keywords: COVID-19, Therapeutic ceiling of care, Prediction model, Variable selection.

Bayesian Inference for Multivariate Spatial Models with R-INLA

Francisco Palmí-Perales[1], *Virgilio Gómez-Rubio*[2], *Roger S Bivand*[3], *Michela Cameletti*[4], *Håvard Rue*[5],

[1]Francisco.Palmi@uv.es, Department d'Economia Aplicada, Universitat de València
[2]Virgilio.Gomez@uclm.com, Departmento de Matemáticas, Universidad de Castilla-La Mancha
[3]Roger.Bivand@nhh.no, Department of Economics, Norwegian School of Economics
[4]Michela.cameletti@unibg.it, Department of Economics, Universitá degli studi di Bergamo
[5]Haavard.rue@kaust.edu.sa, King Abdullah University of Science and Technology

The joint analysis of different variables is continuously increasing in these day and age. Bayesian methods and software for spatial data analysis are generally now well established in the scientific community. However, multivariate spatial analysis can be computationally demanding. Thus, the main aim of this work is to show that R-INLA is a convenient toolbox to analyse different types of multivariate spatial datasets.

Interesting details such as the choice of the prior distribution and the appropriate data structure have been discussed. Additionally, three different datasets have been analysed step by step to illustrate the main goal. The chosen datasets are available in different R packages or can be downloaded from Github. The necessary code to replicate and reproduce the different examples are also available. Finally, it has been shown that R-INLA is a suitable alternative to straightforwardly analyse multivariate spatial datasets.

Keywords: INLA, Bayesian statistics, Multivariate spatial models

Spatio-temporal modeling: a flexible Bayesian approach

Jessica Pavani[1], *Fernando Quintana*[2]

Department of Statistics, Pontificia Universidad Católica de Chile
[1]jlpavani@mat.uc.cl, [2]quintana@mat.uc.cl

In the disease mapping context, data are typically collected for specific regions over time and modeled using parametric spatio-temporal techniques. Within this approach, the spatial dependency is usually accommodated by using a conditional autoregressive prior, whereas the temporal dependency is modeled as either an autoregressive (AR) or a random walk structure. However, the use of these standard methods can provide unsatisfactory results. Ideally, spatio-temporal methods that guarantee greater flexibility to model random effects, mainly regarding spatial dependence, would be desirable. In addition, it would be also of great interest to develop a strategy for defining spatio-temporal clustering. In this circumstance, a Bayesian nonparametric methodology may be used for offering a more coherent modeling framework. This study aims to develop an effective and flexible model to identify and cluster areas where a certain disease behave similarly. Thereby, estimates of relative risk for each area may be provided. To do so, we establish a spatio-temporal model where temporal dependence is defined for areal clusters induced by product partition models (PPM). Unlike similar methods, the PPM produces more flexible clusters, even allowing them to be non-contiguous. To model the temporal component, we define a structure that considers lagged values of observed data, including also a seasonal effect. Our model also considers a spatial effect following a using directed acyclic graph autoregressive, this structure allows the interpretation of a spatial correlation parameter. For the application, we consider the number of cases of dengue, a tropical disease transmitted by mosquito, for all the 145 microregions in the Brazilian Southeast region over an observational period of 12 years, which totals 626 epidemiological weeks. As covariates, we use seasonal indicators and Human Development Index and temporal trend is modeled as an AR(3) plus a seasonal component, one 53-lagged term.

Keywords: Clustering, DAGAR model, PPM model

Love Plot: an illustrative method to show the imbalance between groups.

Judith Peñafiel[1], Natàlia Pallarés[1], Cristian Tebé[1]

[1]jpenafiel@idibell.cat, npallares@idibell.cat, ctebe@idibell.cat, Biostatistics Unit, IDIBELL

BACKGROUND: A common step in observational studies and randomised controlled trials is to assess the balance of relevant covariates between groups. P-values are often used as a measure of between-group imbalance. However, the limitations associated with the (mis)use of p-values are well known. A simple alternative is to compute the standardized mean difference (SMD). For continuous covariates, SMDs provide the same scale and calculation methods are well known. For categorical variables, however, there is less agreement on how to calculate SMDs.

OBJECTIVE: To evaluate the different methods used to calculate balance statistics for a categorical variable.

METHODS: Data were simulated with different assumptions of between-group imbalance in a set of categorical variables and different sample sizes. Common balance statistics for categorical variables were estimated: raw difference (RD) across levels, pooled standardized difference (PSD) across levels and Mahalanobis distance (MD). In addition to the most common ones, the maximum of the RD (MRD) was calculated. The results were presented and compared graphically with the balance statistics on the x-axis and the variable categories on the y-axis, also known as the love plot.

RESULTS: First results suggest that the balance statistics tend to increase as the sample size decreases. However, in most cases these statistics lead to the same conclusions about between-group balance. Statistics across levels, such as RD and PSD, make it possible to observe at which levels the differences are present, but are difficult to visualize graphically. On the other hand, an overall statistic like MD and MRD is easier to visualize graphically. Nonetheless, if there are imbalanced covariates, MD alone may not be sufficient to determine the levels at which differences occur. Similarly, MRD only shows the most unbalanced level. In addition, if a particular level is unbalanced, MD may be overshadowed by the other levels.

CONCLUSIONS: The balance statistics approaches are heuristic approximations, and their exact values are not crucial specially if they are close to 0. A graphical approach, such as love plot, could be an excellent alternative to p-values in assessing imbalance. The choice may depend on the researcher's aims, preference, and the way in which the degree of imbalance between the groups is to be shown.

Keywords: Balance statistics, categorical variables, p-values.

Deep neural learning to predict mRNA editing

Jesús Peñuela[1], Michal Zawisza[2], Carlos Herrera[2], Ferran Reverter[2], Esteban Vegas[2], Jordi García[2]

[1]jpenuela@uoc.edu. Open University of Catalonia
[2]mihizawi@gmail.com, carlos.herrera@me.com,freverter@ub.edu,evegas@ub.edu,jordigarcia@ub.edu.
Department of Genetics, Microbiology and Statistics, University of Barcelona

RNA-editing is a molecular mechanism that performs chemical changes to specific nucleotides in RNA molecules of eukaryote cells and is one of several posttranscriptional mechanisms that adds versatility to the transcriptome. In metazoa, the most usual form of editing is the change from adenine to inosine (A-to-I) performed by proteins of the ADAR family. Although the mechanisms by which ADAR proteins target specific adenosine the regulation of editing activity is not yet fully understood, it has been shown that this regulation plays a key role in several biological processes. In vertebrates, the regulation of specific editing events is an important factor in changes of gene expression during the development of the neural system. Furthermore, certain brain-expressed proteins present very significant differences in editing across different vertebrate clades.

The initial data are genomic sequences annotated with the secondary structures of the pre-mRNA and with the editing positions that were generated by the information tools developed by M. Zawisza. The data generation process is based on three input files: RNA Editing data extracted from REDIPortal, the fasta file of the reference genome and the gene annotation GTF file. After having all the pre-mRNA sequences, the secondary structures are predicted by LinearFold. At the end of this process, the initial data file is generated, which contains all the pre-mRNA sequences along with the prediction of their secondary structure and the annotations of the edited adenosine positions.

The genomic data files had to be processed properly in order to train the deep learning models. For this purpose, we have used fixed length window sequences, centered on each of the adenosines for a given gene. Window sequences of different lengths were tested, finding that the best results were obtained generating windows of $50 + 1 + 50$ nucleotides.

In this work we have used LSTM neural networks with an attention layer. The attention layer is capable of assigning different weights to different positions in each input sequence, seeking to give more relevance to the positions that are more decisive when classifying the sequence.

We analyzed A-to-I RNA editing in the human genome. In balanced scenarios with 800,000 windows sequences, we obtained 93% Accuracy, 93% F1, 0.854 Kappa and 0.95 AUC. We have done preliminary studies in mice and mackerel, and we have achieved 84% and 72% Accuracy, respectively.

Keywords: RNA Editing, LSTM, Attention mechanism.

Assessing the diagnostic ability of medical tests with status defined by right-censored data at a specific time t

Sara Perez-Jaume[1,2], *Jaume Mora*[2], *Josep L Carrasco*[1]

[1]Biostatistics Unit, Department of Basic Clinical Practice, University of Barcelona (SPJ: saraperez@ub.edu, JLC: jlcarrasco@ub.edu)

[2]Sant Joan de Déu Pediatric Cancer Center Barcelona, Institut de Recerca Sant Joan de Déu, Barcelona (SPJ: sara.perez@sjd.es, JM: jaume.mora@sjd.es)

A diagnostic test is a tool applied to individuals with the objective of diagnosing a certain condition or status of health or prognosticating the occurrence of a particular event. Diagnostic tests are usually used to distinguish between diseased and non-diseased subjects and they are of the utmost importance in clinical and biomedical research, since the decisions made on the basis of a diagnostic test might have crucial implications in the management of the patients. The most usual setting considers two possible statuses for the subjects, that is, the outcome of interest is binary. When assessing the diagnostic ability of binary tests, common measures of accuracy are sensitivity and specificity. In the case of continuous diagnostic tests, the area under the receiver operating characteristic (ROC) curve (AUC) is a commonly used index to evaluate the ability of the test to discriminate among the two true states. Moreover, it is usual to choose a threshold that is optimal in some sense to define the test results as positive or negative. One of the most popular criteria to estimate optimal thresholds is the maximisation of the Youden index, which is defined as the sum of sensitivity and specificity diminished in one unit. Once a threshold has been established, a binary test can be constructed from the continuous one and measures such as sensitivity and specificity might also be estimated to assess the diagnostic ability of the test.

Sometimes, the interest lies in predicting the binary status of the subjects at a certain pre-specified time-point t, which is of interest for clinical or biological reasons. When this status is available for all individuals, sensitivity, specificity, AUC and optimal thresholds can be estimated using the standard methodology for the two-state setting. The problem arises when the status is defined by an event occurring at t or before. If the time to event is right-censored (that is, the event time is larger than the censoring time), the status at t is not observed for those patients censored before t. We consider this situation as a missing data problem, and thus we propose two approaches to deal with it. First, we introduce a simple method based on excluding those individuals with missing status at t. Then, we propose a more complex method that uses multiple imputation to recover knowledge about the status of those subjects at time t, using the information provided by the diagnostic test. Those methods are used for estimation and inference about sensitivity, specificity, AUC and optimal thresholds. We illustrate these methodologies by applying them to real case examples about cancer patients. The simulation study demonstrated low biases with both methods and showed the advantages of the imputation method when compared to the missing exclusion approach, especially with small or moderate sample sizes and in scenarios with high proportions of missingness.

Keywords: diagnostic ability assessment, missing data imputation, right-censored data

Mixture cure models with vector and functional covariates

Beatriz Piñeiro-Lamas[1], *Ricardo Cao*[2], *Ana López-Cheda*[3]

[1]b.pineiro.lamas@udc.es, Grupo MODES, Departamento de Matemáticas, CITIC, Universidade da Coruña

[2]ricardo.cao@udc.es, Grupo MODES, Departamento de Matemáticas, CITIC, Universidade da Coruña

[3]ana.lopez.cheda@udc.es, Grupo MODES, Departamento de Matemáticas, CITIC, Universidade da Coruña

Standard survival models assume that, in the absence of censoring, all individuals will experience the event of interest. However, sometimes this is not realistic. For example, if we consider cancer patients being treated and the event is the appearance of an adverse effect, there will be patients that will never experience it. Those who will never develop this health condition will be considered as cured. To incorporate this cure fraction, classical survival analysis has been extended to cure models. In particular, mixture cure models allow to estimate the probability of being cured and the survival function for the uncured subjects. In the literature, nonparametric estimation of both functions is limited to continuous univariate covariates. We fill this important gap by considering both vector and functional covariates and proposing a single-index model for dimension reduction. This approach has been studied in the presence of censoring, but not in the presence of cure. The methodology is applied to a cardiotoxicity dataset from the University Hospital of A Coruña.

Keywords: cardiotoxicity; censored data; survival analysis.

Homicides of social leaders, human rights defenders and signatories of the peace agreement in Colombia 2021-2022 - A first spatial model

John Puerto[1], *Fernando Grajales*[2]

Abstract

In Colombia, the homicide of social leaders has increased significantly after the signing of the peace agreement with the FARC guerrillas, without the apparent recognition and intervention of the Colombian State to reduce this problem. [1] These homicides maintain certain patterns described by different social organizations. [2] This article will analyze the data related to the homicide of Colombian social leaders in the years from 2020 to 2022. To understand these patterns, the application of a statistical methodology that allows us to analyze whether there is spatial dependence in the homicides of social leaders, that is, if there is a pattern and it can be expressed in a grouped, non-grouped (dispersed) or random manner . As a methodology we will use measures of spatial dependence and hypothesis testing, such as Moran's I index. Then a spatial model is proposed with an explanatory variable such as illicit crops in Colombia.

Keywords: Homicides of social leaders Human rights defenders and signatories of the peace agreement in Colombia, Global Spatial Autocorrelation Indices - Moran's I, spatial regression in areas

Referencias

1. Ball, Patrick, and Michael H. Reed. 2016. El Registro y La Medicion de La Criminalidad. El Problema de Los Datos Faltantes y El Uso de La Ciencia Para Producir Estimaciones En Relacion Con El Homicidio En Colombia, Demostrado a Partir de Un Ejemplo: El Departamento de Antioquia (2003-2011). Revista Criminalidad pag: 9 al 23.
2. Ball, Patrick, and Valentina Rozo. 2018. Homicidios de Lideres Sociales En Colombia En 2016 al 2017: Una Estimacion Del Universo. Human Rights Data Analysis Group.
3. Barbona, I. 2014. Ajuste de Modelos Para La Estimacion Espacio y Temporal de Eventos
4. Cliff, A. D., Ord, J. K. 1981 Spatial processes, Pion, p. 21; Bivand RS, Wong DWS 2018 Comparing implementations of global and local indicators of spatial association. TEST, 27(3), 716 al 748 doi:10.1007/s11749-018-0599-x.
5. Comision Interamericana de Derechos Humanos. 2019. Informe Sobre La Situacion de Personas Defensoras de Derechos Humanos y Lideres Sociales En Colombia.

[1]Erazo, Laura Maria Orjuela. 2019. Universidad Jorge Tadeo Lozano The theory of the spiral model applied to the murder of social leaders and human rights defenders in Colombia: between resignation and change.

[2]Latin American Council of Social Sciences. What are the patterns ? Homicides of social leaders in the post-agreement.

To trim or not to trim, that is the question

Pere Puig[1,2]

[1]ppuig@mat.uab.cat, Department of Mathematics, Universitat Autònoma de Barcelona
[2] Centre de Recerca Matemàtica, Campus de la UAB

In clinical settings, the trimmed mean can be used to analyze measurements of vital signs, such as heart rate, respiratory rate, and blood pressure. By removing extreme values caused by stress or other factors, the trimmed mean can provide a more accurate estimate of the patient's condition. Trimmed mean was first documented in an anonymous work in 1821:

...to determine the mean yield of a property of land, there is a custom to observe this yield during twenty consecutive years, to remove the strongest and the weakest yield and then to take one eighteenth the sum of the others. Annales de Mathématiques pures et appliquées, tome 12 (1821-1822), p. 181-204, translated by Huber in 1972.

The α-trimmed mean (TM_α) of the observations is calculated by sorting all the values, discarding $\alpha 100\%$ of the smallest and $\alpha 100\%$ of the largest values, and computing the average of the remaining values. Note that TM_0 is the sample mean and $\text{TM}_{0.5}$ is the sample median, $\text{TM}_{0.25}$ is sometimes called the sample *midmean*. Trimmed means are commonly used in many disciplines and are very useful in machine learning algorithms.

Suppose that we have a data set coming for an arbitrary unimodal symmetric distribution and we want to use an α-trimmed mean to estimate the location parameter μ. How to choose the trimming proportion α? To answer this question we characterize all symmetric distributions with smooth densities such that the α-trimmed mean is an asymptotically efficient location parameter estimator. These are composed of two families of distributions, one of which is unimodal and is referred to as the "H distribution", with density function,

$$h(x-\mu;a,b)=\frac{be^{2b}}{ac(b)}\begin{cases}\exp\left(-\frac{((x-\mu)^2+a^2)b}{a^2}\right) & |x-\mu|\le a\\ \exp\left(-\frac{2b|x-\mu|}{a}\right) & |x-\mu|>a\end{cases}, \tag{1}$$

where $c(b)=(\sqrt{b})\sqrt{\pi b}e^b+1$ and $(.)$ is the error function. The location parameter (population mean) is indicated as μ, and b is a shape parameter that directly determines the truncation proportion α of the trimmed mean from the equation $c(b)=1/(2\alpha)$. This result suggests that an α-trimmed mean will be a good choice for estimating the location parameter μ when the underlying distribution of the data will be similar to the H distribution. Therefore, we propose to fit the data with the H distribution and to estimate the trimming proportion as $\hat{\alpha}=1/(2c(\hat{b}))$ where $\hat{b}$ is the MLE of parameter b. This is an automatic procedure that can be incorporated in a machine learning algorithm. Several exemples of application will be discussed.

Keywords: Asymptotically efficient estimator; Characterization of distributions; Symmetric location models.

Predictability assessment of the first continental heat-cold-health early warning system: new avenues for human health forecasting

Marcos Quijal-Zamorano[1,2], Desislava Petrova[1]; Èrica Martínez-Solanas[1,3], François R. Herrmann[4]; Xavier Rodó[1,5], Jean-Marie Robine[6,7], Hicham Achebak[1,8], Joan Ballester[1]

Affiliations:
1 ISGlobal, Barcelona, Spain
2 Universitat Pompeu Fabra (UPF), Barcelona, Spain
3 Sub-Directorate General of Surveillance and Response to Public Health Emergencies, Public Health Agency of Catalonia, Generalitat of Catalonia, 08005, Barcelona, Spain
4 Division of Geriatrics, Department of Rehabilitation and Geriatrics, Geneva University Hospitals and University of Geneva, Thônex, Switzerland
5 ICREA, Barcelona, Spain
6 Institut National de la Santé et de la Recherche Médicale (INSERM), Montpellier, France
7 École Pratique des Hautes Études, Paris, France
8 Inserm, France Cohortes, Paris, France

The increasing number of extreme climate events due to global warming highlights the urgent need for the implementation of early warning systems directly targeting the effects of weather phenomena on human health. Here we build the first continental heat-cold-health early warning system, and compare its predictability with the original weather forecasts. We did so by considering almost 60 million counts of all-cause mortality in 147 contiguous NUTS regions from 16 European countries, and daily gridded observations and forecasts (with lead times up to 15 days at 24-hour intervals) of 2-meter temperature. We calculated state-of-the-art temperature-lag-mortality models, which account for the delayed effects of daily temperatures on mortality counts. These epidemiological associations were used to transform the daily bias-corrected forecasts into daily predictions of temperature related mortality. We compared the predictive skill of temperature forecasts and temperature related mortality predictions by using predictability assessment techniques widely used in weather and climate forecasting. We additionally quantified the window of predictability by defining the predictability lead time as the lead time in which the predictive skill value falls below specific thresholds. We found that temperature forecasts can be used to issue skillful predictions of heat and cold related mortality accounting for the real impacts of temperature on human health, although the window of predictability was differently reduced by season and location. We also showed that the predictability of the early warnings is to a very large extent constrained by the original weather forecasts, and not by the epidemiological models, which means that further advancements in weather forecasting would automatically turn into an increase in the predictability window of health early warning systems.

Keywords: Early warning systems, bias-correction, predictability assessment, forecasting.

Genetically predicted telomere length and Alzheimer's Disease endophenotypes: a Mendelian Randomization study

Blanca Rodriguez-Fernandez[1], *Natalia Vilor-Tejedor*[2], *Marta Crous-Bou*[3]

[1]brodriguez@barcelonabeta.org, Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation
[2]nvilor@barcelonabeta.org, Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation
[3]mcrous@barcelonabeta.org, Unit of Nutrition and Cancer, Cancer Epidemiology Research Program. Catalan Institute of Oncology (ICO)-Bellvitge Biomedical Research Center (IDIBELL)

Observational studies are designed for measuring an association between an exposure and the occurrence of a disease. However, results may be biased due to confounding factors and the direction of the association may be wrongly determined. Several observational studies have been designed to evaluate the influence of telomere length (TL) on the risk of age-related diseases such as Alzheimer's disease (AD). These studies were limited to concluding whether TL is causally associated with those outcomes. This study aimed to evaluate the potential causal role of TL in AD endophenotypes through a Mendelian randomization (MR) analysis. Our analysis was conducted in the context of the ALFA study. We created episodic memory, executive function, and global cognitive performance composites. We calculated AD and aging signatures as composite measures reflecting cortical thickness of specific AD and aging vulnerable brain regions, respectively. We measured CSF levels of core AD and neurodegeneration biomarkers, that is, amyloid-β ($A\beta$), p-tau, t-tau, and neurofilament light (NfL). A total of 20 single nucleotide polymorphisms associated with TL were used to determine the effect of TL on AD endophenotypes. Analyses were adjusted by age, sex, and years of education. Stratified analyses by *APOE*-$\epsilon 4$ status and polygenic risk score of AD were conducted. Inverse-variance weighted (IVW), maximum likelihood, weighted median and weighted mode methods were used to estimate the causal effect of genetically predicted longer TL on the outcomes of the study. Cochran Q statistic, MR-PRESSO and MR-Egger intercept-test were used as *ad hoc* sensitivity analysis for evaluating the robustness of significant results. Effect sizes were reported in the SD change (β_{IVW}) per copy of the allele associated with longer TL. MR analysis revealed significant associations between genetically predicted longer TL and lower levels of CSF $A\beta$ and higher levels of CSF NfL only in *APOE*-$\epsilon 4$ non-carriers. Moreover, inheriting longer TL was associated with greater cortical thickness in age and AD-related brain signatures and lower levels of CSF p-tau among individuals at a high genetic predisposition to AD. Further observational analyses are warranted to better understand these associations.

Keywords: Alzheimer's disease, causal inference, genome-wide association study

Reference: [1] Rodriguez-Fernandez B., Vilor-Tejedor N., Arenaza-Urquijo E., et al. (2022). Genetically predicted telomere length and Alzheimer's Disease endophenotypes: a Mendelian Randomization study. *Alzheimer's research & therapy*, 14. 167.

COVID-19 patient profiles in the Basque Country: A clustering approach

Lander Rodriguez[1], *Irantzu Barrio*[2], *Daniel Fernández*[3], *José M. Quintana-Lopez*[4]

[1]lrodriguez@bcamath.org, Applied Statistics, Basque Center for Applied Mathematics

[2]irantzu.barrio@ehu.eus, Department of Mathematics, University of the Basque Country

[3]daniel.fernandez.martinez@upc.edu, Department of Statistics and Operations, Universitat Politecnica de Catalunya · BarcelonaTech

[4]josemaria.quintanalopez@osakidetza.eus, Research Unit of the Galdakao-Usansolo University Hospital, Osakidetza Basque Health Service

The ever-increasing availability of medical data has opened the way for Machine Learning and advanced statistical methods to be applied in health. In particular, clustering techniques discover hidden and inherent patterns to organize data into groups without any a priori hypothesis. This feature of clustering techniques can be used for patient profile identification, which would be advantageous for multiple care intervention strategies and to offer an improved medical attention. In this work, we create COVID-19 positives patient profiles from a population-based database based on the novel KAMILA clustering technique (Foss et al., 2016). This technique overcomes the problems faced by clustering methods when dealing with mixed-type data, which is often the case in clinical research. In addition, it is appropriate for large datasets, which is the case in this work.

All the patients included in this study were residents in the Basque Country and were diagnosed COVID-19 from March 1, 2020 until January 9, 2022. The data included sociodemographic data, baseline comorbidities and baseline treatments. In addition, COVID-19 adverse outcomes were also included: hospitalization, adverse evolution (ICU or death) and death. A two-stage process was implemented: first we identified the profiles of COVID-19 positives in the Basque Country and then we assessed their association with the adverse outcomes of the disease. The profiles were created for different periods with the KAMILA clustering technique and their evolution in time was assessed.

Age and the Charlson index were the variables that mainly differentiated the profiles, together with, but to a lesser extent, diabetes, kidney disease, metastatic solid tumor and heart failure. The patient profiles were well differentiated by their risk to the adverse outcomes of COVID-19. Actually, the severity of the outcomes increased with the risk level of the clusters for all the periods and outcomes. Apart from that, the profiles evolved to lower risk profiles along the pandemic, which was reflected in the COVID-19 hospitalization, adverse evolution and death reduction.

To our best knowledge, this is the first study used to create COVID-19 patient profiles from COVID-19 positives of the population and to assess their evolution in time. The previous results suggest the appropriateness of clustering techniques and particularly KAMILA to identify risk profiles in large electronic health records with mixed-type data. This could lead to a better allocation of the health resources and an improved medical attention.

Keywords: COVID-19, clustering.

Directional density-based clustering

Paula Saavedra-Nieves[1], *Martín Fernández-Pérez*[2]

[1]paula.saavedra@usc.es, CITMAga, Universidade de Santiago de Compostela
[2]martin.fernandez.perez0@rai.usc.es, Universidade de Santiago de Compostela

Clustering for directional data has achieved a considerable relevance over the last decades, specially amongst the machine learning community. The most popular approaches are spherical $k-$means with cosine similarity (see [1]) and the use of finite mixture models with von Mises-Fisher components (see [2]). A nonparametric alternative to $k-$means is modal clustering. This approach that associates the notions of cluster and mode, does not require to specify the number of groups in advance (see [3]). The connection between clusters and modes is also present in density-based clustering methodology introduced in [4]. Under this perspective, clusters are identified with the connected components of density level sets. This topic has received remarkable attention in the literature but only for densities supported on an Euclidean space. Concretely, the computational problem of determining the connected components of level sets in high dimensional spaces was addressed in [5] and [6]. As a natural consequence, the empirical mode function and the cluster tree (under the generated hierarchical structure) were defined and an unsupervised classification method was proposed. The main goal of this work is to generalize density-based clustering techniques for directional data. Specifically, we present a novel algorithm for determining the connected components of level sets of densities supported on a unit hypersphere. An extensive simulation study shows the performance of the resulting classification methodology.

Keywords: Density level sets, directional clustering, unsupervised classification.

Acknowledgements. P. Saavedra-Nieves acknowledges the financial support of Ministerio de Ciencia e Innovación of the Spanish government under grants PID2020-118101GBI00 and PID2020-116587GBI00 and ERDF (Grupos de Referencia Competitiva ED431C 2021/24).

1. Bibliography

[1] Dhillon, I. S., and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. Machine learning, 42(1), 143-175.

[2] Banerjee, A., Dhillon, I. S., Ghosh, J., Sra, S., and Ridgeway, G. (2005). Clustering on the unit hypersphere using von mises-fisher distributions. Journal of Machine Learning Research, 6(9).

[3] Oba, S., Kato, K., and Ishii, S. (2005). Multi-scale clustering for gene expression profiling data. In Fifth ieee symposium on bioinformatics and bioengineering (pp. 210-217).

[4] Hartigan, J. A. (1975). Clustering algorithms. John Wiley & Sons, Inc.

[5] Azzalini, A., and Torelli, N. (2007). Clustering via nonparametric density estimation. Statistics and Computing, 17(1), 71-80.

[6] Menardi, G., and Azzalini, A. (2014). An advancement in clustering via nonparametric density estimation. Statistics and Computing, 24(5), 753-767.

A simple procedure for testing the assumption of independent censoring under the mixture cure model when the cure status is partially known

Wende Clarence Safari[1], *Ignacio López-de-Ullibarri*[2], *María Amalia Jácome*[3]

[1]wende.safari@lshtm.ac.uk, Inequalities in Cancer Outcomes Network (ICON), Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom
[2]ignacio.lopezdeullibarri@udc.es, Department of Mathematics, Escuela Politécnica de Ingeniería de Ferrol, University of A Coruña, Ferrol, Spain
[3]maria.amalia.jacome@udc.es, Department of Mathematics, Faculty of Science, University of A Coruña, CITIC, A Coruña, Spain

We develop a simple procedure for testing the independent censoring assumption under the mixture cure model (MCM) in the case that some censored individuals are identified to be cured. This procedure is based on the fact that when the independence assumption between the survival time and the censoring time holds, a MCM-based kernel estimator of the cure probability (Safari et al., 2022) is asymptotically unbiased and consistent. In contrast, the standard regression-based kernel estimator of the cure probability is biased and inconsistent. The regression-based estimator is an extension of the local imputation method in Aerts et al. (2002) to estimate a conditional mean with missingness in the response; the missing at random (MAR) assumption must be valid to apply this estimator. Nonetheless, the MAR condition is not plausible in the MCM with cure status partially known when the independence assumption holds. Consequently, a significant difference between the MCM-based and regression-based estimators is expected, especially when the censoring rate is high. A sensitivity study for the procedure is conducted based on a reasonable range of dependence between survival and censoring times and for different levels of observed cure status. A bootstrap method to approximate the critical values of the testing procedure is proposed.

Keywords: Bootstrap, Censored data, Nadaraya-Watson weights.

References

Aerts M., Claeskens G., Hens N. and Molenberghs, G. (2002). Local multiple imputation. *Biometrika*, 89(2): 375 – 388.

Safari W.C., López-de-Ullibarri I. and Jácome M. A. (2022). Nonparametric kernel estimation of the probability of cure in a mixture cure model when the cure status is partially observed. *Statistical Methods in Medical Research*, 31(11):2164 – 2188.

Disease risk estimation in small areas accounting for local spatial and spatio-temporal discontinuities

Guzmán Santafé[1], *Aritz Adin*[2], *María Dolores Ugarte*[3]

[1]guzman.santafe@unavarra.com, Department of Statistics, Computer Science and Mathematics, Public University of Navarre

[2]aritz.adin@unavarra.es, Department of Statistics, Computer Science and Mathematics, Public University of Navarre

[3]lola@unavarra.es, Department of Statistics, Computer Science and Mathematics, Public University of Navarre

Disease mapping aims to study the spatial and/or spatio-temporal evolution of disease risks or rates. Three main goals are usually pursued: detect high/low risk areas, provide accurate estimates of mortality or incidence risks, and bring to light the spatio-temporal patterns. However, being statistically precise when estimating local disease risk for each area and detecting clusters with high/low-risk areas are somehow contradictory. For clustering detection tasks, methods based on scan statistics are very popular. However, these methods are not suitable for precise risk estimation at area level. By contrast, models with spatial random effects including conditional autoregressive (CAR) priors are used to smooth the risk locally by borrowing information from nearby regions (i.e. neighbor regions in space or space and time), obtaining, therefore, a more stable local risk estimates. Unfortunately, if the disease maps present clusters with high/low-risk areas, smoothing methods based on CAR priors may over-smooth local discontinuities preventing a precise estimation of disease risks.

Previous proposals introduce two-stage approaches to obtain a clustering partition of the areal units (first stage) and to estimate risks by fitting Bayesian hierarchical models including cluster effects (second stage). These proposals rely on hierarchical or density-based clustering methods for the first stage because it was shown that usual clustering methods, such as SatScan, obtained worse results, and the FleXScan method only detects high-risk clusters but not low-risk clusters. Some of these proposals include spatial and spatio-temporal dependence when detecting high-risk and low-risk clusters and when estimating disease risks. However, most of them are limited to model spatial dependence. Additionally, hierarchical and density-based clustering methods do not take into consideration the probability distribution underlying the data (as scan statistics methods do). Given the variability of the observed data, this may lead to non-optimal clustering partitions.

In this talk, we present a new approach based on scan statistics that is able to detect significant spatial and spatio-temporal high/low-risk clusters of areas. This clustering configuration is used in a second stage to improve spatial and spatio-temporal risk estimates. The behavior of the new algorithm is evaluated in a simulation study. It is subsequently used to analyze cancer mortality data in 8000 municipalities of continental Spain.

Keywords: clustering, disease mapping, risk smoothing

Stepped Wedge Randomized Trial: a simulation study.

Naiara Santos[1], *Cristian Tebe*[1]

[1]nsantos@idibell.cat, Biostatistics Unit, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Catalonia, Spain

Background: A stepped wedge trial is a type of crossover trial in which patients are not randomised to one treatment or another, but the groups receive the intervention at different times according to a staggered schedule. The design involves the random and sequential crossover of clusters from control to intervention until all clusters are exposed. The order in which clusters receive treatment is randomised. We will conduct a stepped wedge trial to investigate whether stopping any type of medication intervention for 90 days in terminally ill patients could improve their quality of life without significantly increasing mortality.

Methods: A stepped-wedge, non-superiority, cluster-randomised trial was planned to assess the impact of a tool for adjusting diagnostic and therapeutic intensity compared to traditional clinical practice on the cumulative incidence of 90-day all-cause in-hospital mortality after discharge. Mortality was expected to be 4 in 100 in the control group and 5 in 100 in the intervention group, with a margin of non-superiority for the 5% difference in rates. A generalised linear mixed model with a binomial distribution and log link function was used to estimate relative risk as the measure of effect. We will use the R package swCRTdesign, a package that allows the design and analysis of stepwise studies in an effective way.

Results: The researchers identified 20 eligible doctors and planned to randomise them into 4 sequences and 5 periods, with the first period serving as the baseline. Each doctor would recruit between 20 and 30 subjects at each observed time point. To assess the plausibility of the study and the power of the contrast, a simulation will be performed using the previous assumptions. The simulation results would indicate both the feasibility of the project in terms of the number of patients and doctors to be recruited and the percentage of times that the study hypotheses will be successfully tested.

Conclusions: Stepwise studies are a useful tool for evaluating interventions that cannot be implemented simultaneously in all groups, particularly those that require a training period beforehand. In addition, these types of studies have proven to be more ethical, as all groups receive the intervention at some point in time. Another advantage is that, in addition to allowing comparison between clusters, each group also serves as its own control group.

Keywords: Stepped wedge trial, R package swCRTdesign, Simulation.

When Ecological individual heterogeneity models and large data collide: An Importance Sampling approach

Blanca Sarzo[1,2], *Ruth King*[2], *Víctor Elvira*[2]

[1]Blanca.Sarzo@uv.es, Institut Cavanilles de Biodiversitat i Biologia Evolutiva, University of Valencia
[2]School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh

We consider the challenges that arise when fitting ecological individual heterogeneity models to large data sets. In particular, we focus on (continuous-valued) random effect models commonly used to describe individual heterogeneity present in ecological populations within the context of capture-recapture data, although the approach is more widely applicable to more general latent variable models. Within such models, the associated likelihood is expressible only as an analytically intractable integral. Common techniques for fitting such models to data include, for example, the use of numerical approximations for the integral, or a Bayesian data augmentation approach. However, as the size of the data set increases (i.e. the number of individuals increases), these computational tools may become computationally infeasible. We present an efficient Bayesian model-fitting approach, whereby we initially sample from the posterior distribution of a smaller subsample of the data, before correcting this sample to obtain estimates of the posterior distribution of the full dataset, using an importance sampling approach. We consider several practical issues, including the subsampling mechanism, computational efficiencies (including the ability to parallelise the algorithm) and combining subsampling estimates using multiple subsampled datasets. We initially demonstrate the feasibility (and accuracy) of the approach via simulated data before considering a challenging real dataset of approximately 30,000 guillemots (*Uria aalge*), and, using the proposed algorithm, obtain posterior estimates of the model parameters in substantially reduced computational time compared to the standard Bayesian model-fitting approach.

Keywords: Capture-recapture, intractable likelihood, random effects

Covid-19 in Catalonia: small area estimation of cases, hospitalization, and a causal analysis of the effect of vaccination

Pau Satorra[1], *Cristian Tebé*[2], *Laura Igual*[3]

[1]psatorra@idibell.cat, Biostatistic Unit, Bellvitge Biomedical Research Institute (IDIBELL)

[2]ctebe@idibell.cat, Biostatistic Unit, Bellvitge Biomedical Research Institute (IDIBELL)

[3]ligual@ub.edu, Department of Mathematics and Computer Science, Universitat de Barcelona

Objective: To assess the evolution of COVID-19 cases, hospitalization and vaccination rates in the basic health areas (ABS) of Catalonia using spatial and spatio-temporal modelling techniques, and to perform a causal analysis of the effect of vaccination.

Methods: All data were obtained from open data registries of the Government of Catalonia from March 2020 to July 2022. Primarily, Bayesian hierarchical models, with the Integrated Nested Laplace Approximation (INLA), are used to obtain reliable estimates of cases, hospitalization and vaccination rates in all basic health areas (ABS). The method to perform a casual analysis of the vaccination rates on the hospitalization rates is still to be explored.

Results: For each ABS, on a weekly basis, we obtain different estimates of cases, hospitalizations, and vaccinations: smooth cumulative incidence rates, smooth standardised incidence rates, excess risk and the probability that this excess risk is greater than zero. Furthermore, an R shiny application is under construction to present the evolution of these indicators in a fashionable way through maps and graphical representations.

Conclusions: Small area estimation methods are able to provide more reliable estimates of cases, hospitalization and vaccination rates in each of the ABS, as information about the neighbouring areas and about the past can be used to smooth the results in each area. In addition, this study provides valuable insights into the evolution of the COVID-19 pandemic over time and by small areas. Not only incident cases, but also hospitalization and vaccination rates and how they all interact. Finally, the results are presented in an interactive and user-friendly shiny web application accessible to all public.

Keywords: Small Area Estimation, COVID-19 Surveillance and Causal Analysis

Statistical classification of potential radial glia cells based on nuclear shape measures after mechanical induction from astrocyte cultures

José Pablo Soriano-Esqué[1], *Carlos Borau*[2], *Jesús Asín*[3], *José Manuel García-Aznar*[2], *Soledad Alcántara*[1]

[1]psoriano@ub.edu, Dep. Patologia i Terapèutica Experimental, Institut de Neurociències, Universitat of Barcelona, IDIBELL, Spain

[2]Multiscale in Mechanical & Biological Engineering, Aragón Institute of Engineering Research (I3A), University of Zaragoza, Spain

[3] Dep. Statistical Methods, University of Zaragoza, Spain

Radial glia cells (RG) are the principal embryonic neural stem cell (NSC) that generate different types of progenitors, neurons and glia, also serving as substrate for neuronal migration. At the end of neurogenesis, most RG directly transforms into astrocytes. RG are bipolar cells that form a radial palisade spanning the entire neuroepithelia from the ventricular (apical) to the pial (basal) surfaces. This apical-basal anchorage generates mechanical tensions that are crucial for RG integrity, differentiation potential, and function. This work is part of a research project focused on the contribution of niche mechanobiology in determining RG lineage progression and cell fate.

Primary astrocytes cultures from newborn mice cerebral cortex were grown for 3 days *in vitro* in two class of substrates: polymethyl methacrylate with 2μm linear topographies (ln2PMMA) and borosilicate glass coverslips (control). We have described that the micro3D substrate ln2PMMA mimic surface properties and topology of the RG embryonic niche, signals that are sufficient to induce astrocytes to RG transformation.

We propose a combination of biological, image processing and statistical analysis procedures to unravel the contribution of nuclear deformation induced by ln2PMMA in the mechanical/topological modulation of RG lineage progression.

For that purpose, a MATLAB algorithm has been developed to high-throughput image analysis able to crosslink biological data with nuclear shape measures at a single-cell level. By immunofluorescence and confocal microscopy, we obtained nuclear shape measures (area and eccentricity) of cells identified with specific, cell-lineage markers for astrocytes (GFAP), for RG/NSC (nestin, Pax6 and Sox2), for oligodendrocyte progenitors (NG2) and for tripotential intermediate progenitors (GSX2). Nuclei were stained with Hoechst.

We classified cells/nuclei in the resulting database attending to the lineage markers expressed and according to the literature. Then we define a binary variable that identifies ‘promising’ cells, i.e., cells that would successfully evolve into RG. Similarly, we defined a binary variable for cells that are ‘non-promising’ to be RG. Note that some cells have value 0 in both variables. For each binary variable, we estimated logistic regression models, depending on the substrate, nuclear shape parameters and the cell density. Statistical models estimate the effect of substrate and were interpreted in order to understand the effect of covariates.

Keywords: Neural stem cells, logistic regression, image analysis.

Variable selection strategies applied to identify climatic variables impacting the detection of crop diseases

Suarez F [1], *Fiore J* [2], *Balzarini M* [3], *Gimenez-Pecci MP* [4], *Bruno C* [5]

[1] suarezfranco@agro.unc.edu.ar , Unidad de Fitopatología y Modelización Agrícola (UFyMA) jua– INTA – CONICET

[2] juanmfiore@mi.unc.edu.ar , UFyMA – INTA – CONICET

[3] gimenez.mariadelapaz@inta.gob.ar , Unidad de Fitopatología y Modelización Agrícola (UFyMA). INTA

[4] monica.balzarini@unc.edu.ar , Unidad de Fitopatología y Modelización Agrícola (UFyMA), Estadística y Biometría, Universidad Nacional de Córdoba, Facultad de Ciencias Agropecuarias. Argentina

[5] cebruno@agro.unc.edu.ar , Unidad de Fitopatología y Modelización Agrícola (UFyMA), Estadística y Biometría, Universidad Nacional de Córdoba, Facultad de Ciencias Agropecuarias. Argentina

The epidemiological development of infectious diseases results from the interaction of at least three main factors: a conductive environment (weather conditions), a susceptible host, and a virulent pathogen. Currently, it is possible to easily access large volumes of georeferenced climate data, which is why new models have been generated based on the relationship between environmental conditions and disease in a pathosystem. When it is necessary to model the relationship of the disease with the climate, the challenge of working with multiple climatic parameters appears, generally correlated and redundant. Variable selection is the process of selecting a subset of relevant variables to build robust models. The objective of this work was to compare the performance of three variable selection methods: Stepwise, Boruta (B), and Least Absolute Shrinkage and Selection Operator regression (LASSO) in modelling, context to predict disease risks in crops from high-dimensional climatic variables. Data from three pathosystems, with different *n* records and *p* climatic variables, were processed: one related to Aspergillus flavus, another to Corn stunt spiroplasma (CSS), and the third to Maize dwarf mosaic virus (MDMV), with georeferenced records from the monitoring of these diseases in maize (*Zea mays L.*) crops in Argentina. Each database had the presence/absence value of the pathogen and climatic variables such as relative humidity, temperature, accumulated rainfall, and wind speed, were downloaded, covering the period before sowing until harvest and summarized weekly or monthly. The databases were partitioned into 2 data subsets, one for training (80% of the data) and the other for validation (20% of the data). The selection of variables and the adjustment of the models were carried out with the training base and with repeated cross-validation of k=10 and 5 repetitions. The validation of each model obtained was carried out with the validation base and the selection methods were compared using the values of precision and area under the Receiver Operating Characteristics (ROC) curve obtained. The generated models presented good metrics and inspire the construction of alarm systems for these diseases, based on climatic variables. LASSO produces an intermediate parameterization between the selection methods Stepwise (minimum number of predictors) and B (maximum number of predictors) and achieved greater predictive capacity in the classification models to discern between the presence and absence of the pathogen of the corn. In the case of A. flavus LASSO present a precision of 84.09%, CSS 77.52%, and MDMV 70.69%.

Keywords: LASSO regression, Stepwise, Boruta.

Remdesivir in the treatment of COVID-19: An Inverse Probability of Treatment Weighting analysis

Tebe C[1], *Pallares N*[1], *Langohr K*[2], *Gomez Melis G*[2], *Videla S*[3], *Carratala J*[3]

[1] ctebe@idibell.cat,npallares@idibell.cat , IDIBELL, Barcelona, Catalonia, Spain
[2] lupe.gomez@upc.edu, klaus.langohr@upc.edu, UPC/BarcelonaTech, Barcelona, Spain
[3]svidela@bellvitgehospital.cat, jcarratala@bellvitgehospital.cat, Bellvitge Universitary Hospital, Barcelona, Spain

Introduction: Remdesivir is an antiviral drug approved for the treatment of severe COVID-19. In clinical trials, people who received remdesivir recovered faster than those who received placebo (median 10 vs 15 days), and mortality rates were also improved in those who received supplemental oxygen (4 % vs 13% on day 29 of treatment). There is still debate about the benefits of remdesivir from clinical trials. We used the multicentre DIVINE cohort to assess the effectiveness of remdesivir in COVID-19 hospitalised subjects in three different epidemic waves.

Methods: An inverse probability of treatment weighting (IPTW) propensity score analysis was performed to account for confounding by indication due to the lack of randomisation in treatment assignment in our cohort. Propensity scores were estimated using a logistic regression model stratified by wave. In both cohorts (unweighted and weighted), the incidence of mortality, ICU admission, nephrotoxicity and hepatotoxicity were compared between study groups. A Cox regression model was used to compare the risk of death and a log-binomial model for the other outcomes, resulting in hazard ratio (HR) and relative risk (RR) with a 95% confidence interval. A planned subgroup analysis was performed on identified clinically relevant groups. Analyses were performed using R software (survival, jskm, survey, and lme4).

Results: Among 5813 subjects from the DIVINE cohort, 477 remdesivir users and 1122 non-users were selected. Median age was 64 and 40% were women. During hospitalisation, 36 (7.6%) users and 118 (10.5%) non-users died. After adjustment for IPTW, remdesivir use was not associated with a lower risk of death (HR 0.73 95% CI 0.47 to 1.12), ICU admission (RR 0.84 95% CI 0.58 to 1.22), or nephrotoxicity and/or hepatotoxicity (RR 1.06 95% CI 0.75 to 1.50). Subgroup analysis suggested a potential pretective effect in subjects with early administration of remdesivir. Further research is needed to confirm these findings.

Conclusion: Our preliminary results using real-world data show a crude protective effect of remdesivir on in-hospital mortality, but the effect is minimised after adjustment for key confounders. No safety concerns with regards to renal and liver outcomes were raised in subjects with COVID-19 treated with remdesivir in our cohort. Further methodological approaches are planned to confirm these results.

Keywords: real world data, propensity score, inverse probability weighting, cox regression, log-binomial.

Partial Least Squares for binary data and its associated biplot applied to the classification of Colletotrichum Graminicola strains

Laura Vicente-Gonzalez[1], *Jose Luis Vicente-Villardon*[2]

[1]laura20vg@usal.es, Departmento de Estadística, Universidad de Salamanca
[2]villardon@usal.es, Departmento de Estadística, Universidad de Salamanca

In this work we propose a generalization of Partial Least Squares Regression where all the variables, responses and predictors, are binary. The method is named Binary Partial Least Squares (BPLS). A representation for BPLS, that combines two logistic biplots for responses and predictors, is also described.

The final algorithm is based on a generalization of NIPALS to handle binary variables, extending also a procedure recently proposed by the authors.

The method is applied to the classification of several strains of Colletotrichum Graminicola using RNA data. The differences among nine strains, corresponding to their countries of origin, and the genes that characterize them are studied.

For the calculations we have used the `R` software. New functions have been included in the package `MultBiplotR`.

Keywords: Binary Data, PLS, Biplot

How do past training exposures affect injury risk in football?

Lore Zumeta-Olaskoaga[1,2], *Andreas Bender*[3], *Dae-Jin Lee*[4]

[1]lzumeta@bcamath.org, BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
[2]Departamento de Matemáticas, Universidad del País Vasco UPV/EHU, Leioa, Spain
[3]andreas.bender@stat.uni-muenchen.de, Statistical Consulting Unit StaBLab, Ludwig-Maximilians Universität München, Munich, Germany
[4]daejin.lee@ie.edu, School of Science and Technology, IE University, Madrid, Spain

The study of training load and its influence on the risk of sports injuries is one of the hot topics in sports injury research, including in the specific field of football. Gaining more insights into the area of training load empowers coaches, physiotherapists, sports physicians, and other professionals involved in a football club to make the most of this knowledge, such as: developing effective training plan strategies that will enhance players' performance while also lowering their risk to injury.

To model the effects of past training exposures on the risk of sports injuries, the time-varying nature of both, exposure and the outcome variable, should be considered. Besides, the effects of past exposures may cumulate over time and may present complex forms of association. And as a player may sustain more than one injury, dependencies induced by these subsequent injuries should also be taken into account. In this regard, we propose the use of piece-wise exponential additive mixed models for modelling such data. In this work in particular, we study time-varying exposures as weighted cumulative effects, i.e. assigning weights to past training exposures based on the time elapsed since the exposure occurred.

We conduct a simulation study to evaluate the performance of the proposed model, under different true weight functions and different levels of heterogeneity between recurrent events. Finally, we illustrate this recurrent events flexible modelling approach with the application of a case study of an elite male football team participating in LaLiga. The cohort includes time-loss injuries as well as external training load (e.g. training and competition time, power output, distance, sprints, speed etc.) tracked by Global Positioning System devices, during the seasons 2017-2018 and 2018-2019.

Keywords: piece-wise exponential additive mixed models, recurrent events, sports medicine.

Pósteres

Bayesian estimation of transition probabilities in multi-state models: study of hospitalization of severe influenza cases

Lesly Acosta[1], *Carmen Armero*[2]

[1]lesly.acosta@upc.edu, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona-TECH

[2] Departament d'Estadística i Investigació Operativa, Universitat de València, Spain.

Multi-state models are survival models that study the transit of individuals in a target population between different states over time. In medical contexts, states often represent different situations of illness and/or health. Outcomes of interest in these models are times between transitions and trajectories through the states of the process. Our work focuses on the second issue raised.

Transition probabilities between states are conditional probabilities generated by binomial or multinomial sampling models. The Bayesian approach is adopted to estimate the posterior distribution for these conditional probabilities using conjugate Beta-Binomial and Dirichlet-Multinomial models. In addition, we account for the posterior distribution of the unconditional probabilities associated to the different absorbing states of the model by means of simulation tools.

Data for the analysis were collected from a retrospective cohort study of 1306 hospitalized laboratory confirmed influenza (SHLCI) patients registered by the 14 hospitals included in the Influenza Surveillance System of Catalonia (PIDIRAC) from 1 October 2017 to 22 May 2018. All patients were initially attended by a physician at admission registration, and then were redirected to either Ward 1 or to ICU. Patients on ward 1 could be sent to ICU, discharged home, derived to a long-term care facility or die. Patients in ICU can die or, if they improve, be sent to a second ward of the hospital, from where they can die, be discharged home or be sent to a long term facility center.

The results of the combined use of the Bayesian approach within a multi-state model framework are advantageous. The posterior distributions obtained would contain all relevant information over the transition probabilities of interest, and thus would also allow to gain a better insight of the clinical evolution of the influenza disease. Overall, they may be a useful tool in the effective management of Influenza hospitalized patients during peak influenza epidemic activity.

Keywords: Conditional probabilities; Confirmed influenza hospitalization; Survival analysis.

Assessing risk factors of evolutive health-related quality-of-life and mortality: a joint modelling approach

Urko Aguirre[1,2,3], Adrian Lopez[4], Marta Jiménez Toscano[4], Jose María Quintana[1,2,3]Maximino Redondo[3,5] , for the REDISSEC-CARESS/CCR group

[1]urko.aguirrelarracoechea@osakidetza.eus, Research Unit, Osakidetza Basque Health Service, Barrualde-Galdakao Integrated Health Organization, Galdakao-Usansolo Hospital Galdakao, Spain

[2]Kronikgune Institute for Health Services Research, Baracaldo, Spain

[3]Network for Research on Chronicity, Primary Care, and Health Promotion (RICAPPS)

[4]Department of Surgery, Hospital del Mar, 08003 Barcelona, Spain.

[5]Unidad de Investigación, Hospital Costa del Sol, Malaga, Spain. CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain

Introduction and Objective(s): Nowadays, colorectal cancer (CRC) is a major socio-health problem, being one of the most frequent causes of death in the general population. Although the incidence of CRC has increased slightly in recent years, patient survival has stabilised, mainly due to advances in both diagnosis and treatment. However, few clinical care guidelines exist to adequately address deficits in the Health-Related Quality of Life (HRQoL) of these patients. In longitudinal studies, information on HRQoL is observed in different measurements until the event of interest (e.g. mortality) occurs. The aim of this work is to study how changes in HRQoL vary over time and influence mortality in patients with CRC, simultaneously assessing the association between both outcome variables. Furthermore, it is in our best interest to identify which patient sociodemographic and clinical features are related to this mechanism.

Method(s) and Results: The subjects of the present study are patients with colorectal cancer and belonging to the CARESS-CCR cohort (22 hospitals, 5 Autonomous Communities). Information was collected on sociodemographic and clinical data. Regarding HRQoL, these patients completed the EORTC QLQC30 questionnaire, in 6 different measurements, at baseline point and at follow-up up to 5 years (1 month, 1-2-3 and 5 years). Joint models were developed in two steps: first, generalized mixed models were applied to predict the evolution of the summary component of the EORTC QLQC-30, as well as survival models for mortality. Finally, the association between the two modelling approaches was assessed. Results of the joint model showed that the Charlson comorbidity index was associated with worse HRQoL outcomes and higher mortality risk. An improved HRQoL outcome was associated with a lower 5-year mortality risk ($p<0.001$).

Conclusions: Joint models are a useful statistical method to characterize HRQoL trajectories in the first 5-years post-surgery in CRC. A better evolution of HRQoL is associated with a less risk of mortality at 5 years post-surgery.

Keywords: HRQoL, joint modelling, mortality, colon cancer.

Impact of the evolution of the sea surface temperature on the coasts of the Valencia Region

Laura Aixalà-Perelló[1], *Xavier Barber*[2], *Antonio López-Quílez*[3]

[1]laixala@umh.es, Centro de Investigación Operativa, Universidad Miguel Hernández
[2]xbarber@umh.es, Centro de Investigación Operativa, Universidad Miguel Hernández
[3]antonio.lopez@uv.es, Departamento de Estadística e Investigación Operativa, Universidad de Valencia

Many sectors that rely heavily on natural ecosystem conditions, such as marine aquaculture, are affected by environmental instability that may be associated with climate change. In order to optimally plan the development of fish or shellfish farming, it is crucial to incorporate climate risk as a decisive factor in selecting marine areas that will henceforth maintain suitable conditions.

Sea surface temperature (SST) is an example of an unstable phenomenon relevant to the fishing industry. In the literature there are many approaches to accurately define and predict the behavior of extreme events caused by environmental instability.

Specifically, there are four lines of research in the framework of spatio-temporal modelling of extreme phenomena. Firstly, conditional models of a spatio-temporal process given a point in time. Secondly, dynamic time series models that integrate spatial dependence. Moreover, hierarchical models in which spatial dependence is incorporated through the random components of the models. And finally, but widely applied, maximum stability processes that represent the interactions of space and time.

SST instability has been analysed by applying two techniques for different purposes. We started by modelling the data using dynamic time series models to assess the temporal behaviour, and then hierarchical models were applied to look at the behaviour of the region as a whole. Finally, the impact of the thermal instability of the sea was studied with the maximum stability process.

Inference and predictions on hierarchical models and maximum stable processes are approximated using the R-INLA package. The results of each analysis provide us with an estimate of the most stable marine areas where fish or shellfish farming will maintain optimal conditions for development.

Keywords: Extreme events, Spatio-temporal models, R-INLA.

A zero-inflated Bayesian modeling of sports injury risk incidences

Oihane Álvarez[1], *Lore Zumeta-Olaskoaga*[1], *Joaquín Martínez-Minaya*[2], *Dae-Jin Lee*[3],

[1]{oalvarez,lzumeta}@bcamath.org, Applied Statistics Research Line, BCAM - Basque Center for Applied Mathematics, Bilbao, Bizkaia, Spain
[2]jmarmin@eio.upv.es, Department of Applied Statistics and Operations Research and Quality, Universitat Politècnica de València (UPV), Valencia, Spain
[3]daejin.lee@ie.edu, School of Science and Technology, IE University, Madrid, Spain

Sports injury risk incidence is a significant concern for athletes, coaches, and medical professionals in any sport or physical activity. Injuries can not only limit the ability to engage in sports, but also have long-term negative effects on health. Therefore, it is crucial to take preventive measures and establish risk management strategies to reduce the incidence of injuries. Hence, accurately modelling injury risk can aid in injury prevention and management strategies. One statistical method used for modeling count data is the Poisson regression model. However, in many cases, the count data may have an excess of zeros, which violates the assumptions of the Poisson regression model. The zero-inflated Poisson (ZIP) model is a commonly used approach to handle such data, accounting for both the overdispersion and excess zeros.

Bayesian methods provide a flexible framework for modeling complex data structures and incorporating prior knowledge. Moreover, Bayesian inference allows for uncertainty quantification, which we found particularly useful in the context of sports injury risk incidence. Recently, some studies have shown a relationship between maturity and growth timing and sports injuries. There exists some factors that may increase the risk of sports injuries in individuals who are still growing and developing. For instance, during periods of rapid growth, bones, muscles, and tendons may be more vulnerable to injury. This is because growth can occur at different rates, causing the bones and muscles to grow at different rates, which can lead to imbalances and increased stress on the body.

In this work, we illustrate a zero-inflated Poisson and Binomial likelihood model, where we allow for a linear predictor in both the zero-inflation and in the mean. We analyzed data from a study conducted on 110 football players from a professional youth academy team over two decades. Multiple predictive factors which allows exploring the possible influence of different variables on injury burden such as maturity or timing of injury were included in both linear predictors: i) the one corresponding to the zero-inflated part, and ii) the one corresponding to the count part. More robust and reliable results than frequentist ZIP models were obtained. Our results show that the proposed method can provide valuable insights for injury prevention strategies in sports.

Keywords: Zero-inflated Poisson, Bayesian inference, Sports injuries risks incidences.

Is my vagina stressed? Bayesian Dirichlet models to investigate the effect of stress on vaginal microbioma in a Spanish cohort

Rubén Amorós[1], *Joaquín Martínez-Minaya*[2], *Blanca Sarzo*[3,4], *Rima Abumallouh*[5,6], *Giuseppe D'Auria*[7], *Natalia Marin*[6,8], *Raúl Beneyto*[7], *Maria-Jose Lopez-Espinosa*[5,6,7,8]

[1]ruben.amoros@uv.es, Departament d'Estadística i Investigació Operativa, Universitat de València
[2]jmarmin@eio.upv.es, Department of Applied Statistics and OR, and Quality, Universitat Politécnica de Valencia
[3]Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València
[4]School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh
[5]Epidemiology and Environmental Health Joint Research Unit, FISABIO-Universitat Jaume I-Universitat de València
[6]Spanish Consortium for Research on Epidemiology and Public Health (CIBERESP, Madrid)
[7]FISABIO-Public Health (Valencia)
[8]Faculty of Nursing and Chiropody, Universitat de València.

The vagina is inhabited by numerous microorganisms which constitute the vaginal microbiota. Its composition is dynamic and evolves throughout a woman's life. Previous studies on rodents have shown that stress could play an important role in changes in the vaginal microbiota, by altering cortisol levels which, by inhibiting the deposition of estrogen-dependent substrates, would limit the growth of *Lactobacillus*. Such an imbalance in the vaginal microbial community predisposes women to a higher risk of suffering from sexually transmitted diseases and other gynaecological problems.

However, the examination of these relations presents certain challenges, as microbiota data consists of the proportions of genetic sequences for several hundreds of genus or species. As a result, when analysing this kind of high-dimensional compositional data (CoDa), specifically tailored statistical techniques for dimensionality reduction shall be considered.

In this work, we present the preliminary results of the possible association between vaginal microbiota and stress, quantified as the concentration of cortisol in blood serum and hair on 259 women from INMA (Environment and Childhood) Project. Dirichlet multinomial mixture models were used to reduce the dimensionality of the vaginal data into the probability for each woman to belong to one of few microbiota profiles. The impact of cortisol levels on the microbiome profile, corrected by other biological and socioeconomic variables, is then assessed in a Dirichlet Bayesian regression model.

Keywords: Microbiota, compositional data, hierarchical Bayesian models

Funding: AECC-Seed Ideas 2019 (IDEAS19098LOPE) and Generalitat Valenciana (CIGE/2021/071) and (AICO/2021/182).

Impact of the environment on health status of intensive care unit patients: functional data analysis using wearable monitoring systems

Juan A. Arias-Lopez[1,2], *Carmen Cadarso-Suarez*[1,2], *Manuel Oviedo de la Fuente*[3], *Pablo Jesus Lopez-Soto*[4,5,6]

[1]juanantonio.arias.lopez@usc.es, Biomedical Data Science Unit, Department of Statistics, Mathematical Analysis, and Operational Research, Universidade de Santiago de Compostela
[2]CITMAga, 15782 Santiago de Compostela
[3] CITIC, Research Group MODES, Department of Mathematics, Universidade da CoruÃ±a
[4]Comprehesive Nursing Care. Multidisciplinary Perspective. Instituto Maimonides de Investigacion Biomedica de Cordoba (IMIBIC)
[5]Department of Nursing, Hospital Universitario Reina Sofia de Cordoba
[6]Department of Nursing, Pharmacology and Physiotherapy, Universidad de Cordoba

The aim of this study was to identify the environmental variables in a critical care unit (ICU) that have the most significant impact on patient health and to determine whether functional data analysis (FDA) is a useful method for analyzing continuous data obtained from wearable devices. Physiological and environmental data were collected from 77 patients using wearable devices through an IoT platform.

FDA is a field in modern statistical science that allows to analyse complex data objects (e.g. vectors, curves, images ...) that provide more information than traditional multivariate analysis. We used FDA tools to identify the environmental variables that were most strongly correlated with physiological variables, and tested linear and non-linear regression models to predict patient health outcomes.

Our analysis revealed that the average noise intensity in the ICU was the most important variable, together with the previous 15-minute interval functional curve for a given health outcome. These results show that ICU noise levels have strong impacts on cardiac frequency, systolic blood pressure, and diastolic blood pressure. Our results also suggest that FDA is a useful method for analysing continuous data from wearable devices and that interventions to reduce noise levels in the ICU may have significant positive effects on patient health outcomes.

Keywords: Functional Data Analysis, critical care, wearable technology

Genome simulation study in GWAS models with heteroscedasticity across management regimens

Eugenia Bortolotto[1], *Cecilia Bruno*[2,3]

[1]bortolotto.eugenia@gmail.com, Doctorado en Estadística. Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario (UNR), Rosario, Argentina

[2]cebruno@agro.unc.edu.ar, Estadística y Biometría. Facultad de Ciencias Agropecuarias. Universidad Nacional de Córdoba (UNC). Ciudad Universitaria, Córdoba, Argentina

[3]cebruno@agro.unc.edu.ar, Unidad de Fitopatología y Modelización Agrícola (UFYMA - CONICET), Córdoba, Argentina

With the goal of evaluate association mapping (AM) models to detect molecular markers significant in cases where the phenotype is measure in more than one environment and the homoscedastic model cannot assumed, we compared Genome-Wide Association Studies (GWAS) models with different levels of variance. In this study, a dataset based on vegetal parameters was simulated in *xbreed* package in R. There is a database of 240 individuals, 79918 SNP-type (Single-Nucleotide Polymorphisms) molecular markers and 15 QTL. A phenotype trait was also simulated with normal distribution in each environment. Then, to get heteroscedasticity, we simulated three phenotypic repetitions for each genotype at each environment (E1 and E2) by adding a random error term normally distributed with zero mean and three different variances across the environment. First setting with σ^2=0.25 in both environment; second setting 0.25 and 2.25 for σ^2 in E1 and E2, respectively; and the last one setting 0.25 and 12.25 for σ^2. To control for heteroscedasticity, we fitted a Linear Mixed Model (LMM) with Environment (E) as fixed effect and Genotype (G) and G×E interaction (GE) as random effects. In addition, several variance structures were tested in each case and compare with Akaike and Bayesian Information Criteria (AIC and BIC, respectively). Then, the Best Linear Unbiased Predictors (BLUPs) of the model are use in the GWAS. The seven AM models evaluated ranged from simple to complex and included: General Linear Model with Principal Component Analysis, GLM-PCA; Linear Mixed Model with PCA+K (Kinship matrix for family relatedness estimates), LMM-PK; Compressed LMM, CLMM; Enriched CLMM, ECLMM; settlement of LMM under progressively exclusive relationship, SUPER; multiple loci LMM, MLMM; and fixed and random model circulating probability unification, FarmCPU; all of these models were fitted in *GAPIT* package in R. We also compare this model with the LMM considering the markers as covariance matrix with *Sommer* package in R. The correlated multiple testing was done by Li and Ji method that is based in eigenvalues of a correlation matrix. We examine Quantile-Quantile (Q-Q) plots to determining if models control false positives and false negatives. Then compare the number of significant markers identified by Li and Ji method in eight different association models and analyse the proportion of significant SNPs that are less than 5cM from a simulated QTL. The MLMM and ECLMM models were the ones that detected true associations at the different levels of heteroscedasticity. The number of false negative is much higher with the GLM-P and FarmCPU models.

Keywords: variance components; molecular markers; genotype-by-environment interaction, variance heterogeneity, best linear unbiased predictor

Statistical techniques and software used in the field of Clinical Medicine: a bibliographic review

Gerard Boxo[1], *Rosa Abellana*[1], *Josep L Carrasco*[1]

[1]jlcarrasco@ub.edu, Biostatistics, Department of Basic Clinical Practice, University of Barcelona

In the Health Sciences courses, students are typically introduced to Statistics through a basic core subject. The course syllabus varies according to the number of ECTS in each course. However, regardless of the teaching time, the syllabus includes statistical inference techniques, hypothesis contrasts testing and modeling approaches. Often, the statistical techniques presented in these subjects are heirs to teaching plans defined years ago.

Additionally, among the different teaching activities included in these subjects, we find the computer laboratories, in which students get in contact with statistical data analysis. This activity involves the use of statistical software that allows data management and the execution of appropriate statistical analyses. Currently, the range of statistical programs is very wide and the teacher must decide which software will be chosen. The factors for making this decision can be several. Examples are the cost, the teacher's familiarity with the software or the students' access to the software outside the classroom.

The objective of this work is to evaluate which are the statistical techniques, as well as the statistical software, that researchers in the field of Clinical Medicine are currently using. Thus, it will be possible to assess if the syllabus of Statistics basic subjects concurs with the methodology that the future healt professionals will meet in their research practice.

To achieve the objective proposed here, a bibliographic search will be carried out. Research papers published in 2022 will be downloaded by means of web scrapping using the Selenium Python's package. The journals considered are The Lancet, The New England Journal of Medicine, the Journal of the American Statistical Association and the British Medical Journal. These are the four top rated journals in the General Medicine category according to their impact factor. The information about statistical techniques, software and type of research design (experimental or observational) will be extracted from the downloaded papers using text mining techniques. Specifically, an algorithm of unsupervised autonomous learning will be applied.

Keywords: Text mining, Clinical Medicine, Statistical techniques, Statistical software

Factor analytic biplot in multienvironment trials with incomplete data

Cecilia Bruno[1], *Mónica Balzarini*[2]

[1]cebruno@agro.unc.edu.ar, Unidad de Fitopatología y Modelización Agrícola (UFyMA). INTA-CONICET. Estadística y Biometría. Facultas de Ciencias Agropecuarias. Universidad Nacional de Córdoba. Argentina.

[2]monica.balzarini@unc.edu.ar, Unidad de Fitopatología y Modelización Agrícola (UFyMA). INTA-CONICET. Estadística y Biometría. Facultas de Ciencias Agropecuarias. Universidad Nacional de Córdoba. Argentina.

In multi-environmental trials, the additive main effects and multiplicative interaction models (AMMI) are used to explore the genotype by environment interaction (GEI) based on a complete dataset where all genotypes were evaluated in all sites. The biplot from principal component analysis of residuals from the additive model (AMMI-biplot) allows us to visualize the ordination of the genotypes (G) according to their performance across environments (E), *i.e*., locations and years combination. In incomplete datasets, the linear mixed models (LMM) using factor analytic (FA) covariance structure arise as an alternative to predict the GEI effect. Then, an FA-biplot could be obtained by plotting the G and E scores derived from the FA-LMM. The aims of this work were: 1) to quantify the consensus about G ordination achieved from the AMMI-biplot with the ordinations of G built from the FA-biplot obtained from complete multi-location and multi-year testing by the linear mixed model using FA covariance structure, and 2) to assess the impact of the increasing levels of missing G, discarded at a third year of testing due to low performances, on the final G ordinations. The G ordinations were assessed on four datasets from wheat testing with different arrangements of the number of G and E: 10G×15E, 15G×15E, 10G×30E, and 15G×30E. We started with complete datasets where all G had been evaluated in all E for three consecutive years. Then, to generate different levels of incompleteness, we eliminated the G with the worst performance in the last year of evaluation. Thus, we eliminated G simulating the selection of G in the Argentinean wheat plant breeding programs until we reached 50% of the missing genotypes in each database. To measure consensus we used a Generalized Procrustes Analysis. Despite applying two different procedures to obtain the biplots, one based on a fixed effects AMMI model and the other based on a random effects FA model, the biplots showed the same GEI pattern. With increasing levels of missing G, the consensus between the interaction patterns was also statistically significant demonstrating that FA-biplots are robust tools for ordering G under the studied mechanism of missing data.

Keywords: Genotype by environment interaction, Additive main effects multiplicative interaction (AMMI), Factor analytic covariance structure.

Proposal of statistical matching methodology to fuse data from different survey samples

Naroa Burreso Pardo[1], *Edorta Arana Arrieta*[1], *Libe Mimenza Castillo*[1], *Irantzu Barrio*[2,3], *Josu Amezaga Albizu*[1]

[1]naroa.burreso@ehu.eus, Department of Audio-visual Communication and Advertising, University of the Basque Country UPV-EHU

[2]Department of Mathematics, University of the Basque Country UPV-EHU

[3]BCAM, Basque Center for Applied Mathematics

Statistical matching (SM) refers to a series of methods that use different available data sources (usually survey samples), referred to the same target population, with the aim of studying the relationship among variables not jointly observed in a single data source. In the simplest case of statistical matching there are two different databases A and B, each of them with their specific set of variables **X** and **Y** respectively, and they also share a set of common variables **Z**. The objective of the fusion is to investigate the relationship between **X** and **Y**. To study this relationship different methods exist. Most of the SM techniques assume that A and B are random samples of independent and identically distributed observations selected from the same infinite population. However, in most real cases the data come from surveys based on some complex design carried out on the same finite population. One of the characteristics of this type of data is the sampling weights, which indicate the number of units that each sampled observation represents in the finite population.

The goal of this methodological proposal is to use SM in order to create a synthetic database powered by information from two different surveys conducted to the same finite population, where each survey may (or may not) be based on a different sample design. To do so, one of the databases will be considered as the donor (the one with more individuals) and the other one as the receptor (the one with less individuals). The donor will transfer information to the receptor. The proposed methodology consists of five main steps: 1) Identify the common stratification variables and create the donation clases using those variables; 2) find the common variables (matching variables) not used in the sampling design; 3) calculate the distances between the individuals (in the common variables and same donation clases) using Gower´s distance (Gower 1971); 4) establish the donor- receptor relations taking into account the distances among individuals and their sampling weights; and 5) transfer the information from the donor database to the receptors using the relations established before. Once the fusion is completed, two types of analysis will be carried out to check the validity of the synthetic database: the imputed variables conserve the marginal distribution and the imputed variables conserve the joint distribution with the matching variables.

The proposed methodology has been applied to the databases Ikusiker and CIES. Ikusiker is a panel that measures audience consumption in non-traditional media of students between 12 and 21 years old from the four provinces in Southern Basque Country (Araba, Bizkaia, Gipuzkoa and Nafarroa). CIES is a survey that measures audience consumption in traditional media of the population over 14 years old in the same four provinces. The final product of the fusion has been a complete database with the information of the variables available in both original databases.

Keywords: Statistical matching, surveys, design based inference.

A Bayesian Gompertz approach to evaluate the optimal surgical space for laparoscopic surgery

Gabriel Calvo[1], *Carmen Armero*[1], *Virgilio Gómez-Rubio*[2], *Guido Mazzinari*[1, 3]

[1]gabriel.calvo@uv.es, Department of Statistics and Operations Research, Universitat de València

[2]Departamento de Matemáticas, Escuela Técnica Superior de Ingenieros Industriales, Universidad de Castilla-La Mancha

[3]Research Group in Perioperative Medicine and Department of Anaesthesiology, Hospital Universitari i Politècnic la Fe

Laparoscopy is a surgical procedure performed using small incisions and a camera to visualise organs or conduct minor surgeries in the abdomen or pelvis. To create enough space for the surgical instruments, carbon dioxide (CO2) is insufflated into the abdomen. It is crucial to identify the critical point at which insufflation should be limited to maximise surgical space and minimise harmful effects.

To evaluate the relationship between insufflation pressure and intra-abdominal volume generated and make inferences for some interesting outcomes of the procedure, a Bayesian logistic growth curve was used in previous studies. In this work, a Bayesian Gompertz growth mixed-effects model is proposed. The asymptotic deceleration point of the Gompertz curve, calculated from the fourth derivative of the function, can be considered as the critical point beyond which it is expected that the increase of the surgical space is not of practical interest. The main goal of this study is to compare the two-modelling approaches and select the option that best suits the relationship between insufflation pressure and intra-abdominal volume generated in a laparoscopy.

The data upon which our model has been applied were obtained from 49 patients who underwent laparoscopy at the Hospital Universitario y Politécnico La Fe (València) between March 2021 and January 2023. For each patient, the intra-abdominal volume was recorded based on the insufflated intra-abdominal pressure at the start of surgery. Furthermore, variables including age, sex, weight, height, number of previous pregnancies, and several anthropometric measures, such as sagittal abdominal diameter, body mass index, and conicity index, were also recorded.

Keywords: Critical growth points; Non-linear mixed models; Repeated measures.

Managing REDCap Data: The R package REDCapDM

João Carmezim[1], *Pau Satorra*[1], *Judith Peñafiel*[1], *Esther García-Lerma*[1], *Natalia Pallarés*[1], *Naiara Santos*[1], *Cristian Tebé*[1]

[1] jcarmezim@idibell.cat, psatorra@idibell.cat, jpenafiel@idibell.cat, egarcia@idibell.cat, npallares@idibell.cat, ctebe@idibell.cat, Biostatistics Unit, Bellvitge Biomedical Research Institute (IDIBELL)

BACKGROUND: REDCap (Research Electronic Data CAPture) is a web application for creating and managing online surveys and databases. Clinical data management is essential before performing any statistical analysis to ensure the quality and reliability of study data.

OBJECTIVES: The aim of the R package "REDCapDM" is to process REDCap data and provide useful tools to perform all tasks involved in the data cleansing process prior to statistical analysis.

METHODS: The 'REDCapDM' package is organized into four dimensions, each serving a specific purpose. First, reading and processing raw data from REDCap or through a REDCap API connection in R. Second, performing data transformation and data organization. Third, it identifies queries, specifically missing values, values that fall outside the lower and upper limit of a variable and other types of inconsistencies in data from REDCap in R. Fourth, it performs an automatic control of queries that have been resolved or are pending resolution.

RESULTS: A total of 5 functions have been included in the package to speed up and automate the process of data managing of REDCap data in R. The "redcap_data" function is used to read and prepare the data set originated from REDCap. The "rd_transform" function is used to perform several data transformations. The "rd_query" and "rd_event" functions allow the identification of different types of data issues and generate a query report. This query report is consists of two elements: a data frame containing the information needed to identify each query in the REDCap project and a summary of the number of queries per variable/event and query type. Finally, we have the "check_queries" function, which is used to compare a previous dataset of queries with a new one and check which queries are new, which are pending resolution, which are miscorrected and which are solved.

CONCLUSIONS: The R package "REDCapDM" contains a collection of useful functions to import, organize, and perform a quality-checking of data from REDCap. This package fills a gap in the available tools for managing REDCap data, making it an invaluable asset for researchers. The "REDCapDM" package is freely available from the CRAN library (https://cran.r-project.org/web/packages/REDCapDM/index.html) and is updated regularly.

Keywords: REDCap, clinical data management, query.

Comparison of Statistical Approaches for Interval-Censored Data: Analysis of data from an HIV-negative MSM Cohort

Inês Carvalho[1], *Leandro Duarte* [2], *Carla Moreira* [2,4,5], *Luís Machado* [2], *Ana Paula Amorim* [2], *Miguel Rocha* [3], *Paula Meireles* [4,5]

[1]b13232@math.uminho.pt, Centro de Matemática da Universidade do Minho, Universidade do Minho, 4800-058 Guimarães, Portugal

[2] Centro de Matemática da Universidade do Minho, Universidade do Minho, 4800-058 Guimarães, Portugal

[3] Grupo de Activistas em Tratamentos, Lisboa, Portugal

[4] EPIUnit - Instituto de Saúde Pública, Universidade do Porto, Rua das Taipas, n° 135, 4050-600 Porto, Portugal

[5] Laboratório para a Investigação Integrativa e Translacional em Saúde Populacional (ITR), Universidade do Porto, Rua das Taipas, n° 135, 4050-600 Porto, Portugal

Survival analysis differentiates from other domains in statistics as the data can be censored, meaning that the event during a follow-up period, such as death or occurrence of a disease, is incomplete. In standard survival analysis, the event of interest might be observed exactly or is right-censored, that is, individuals will be event-free througout follow-up, but the event may occur after the last observation time. In other situations, however, the times of the events of interest are known only to have occured within a time interval from the last examination without the event to the first examination after the event has occured. These data are known to be interval-censored. Although interval-censored data requires specific methods, those developed for right-censored data are still standard practice. A common ad-hoc approach for dealing with interval-censored data is to assume that the event has occurred at the end, midpont or beginning of each interval and then apply the usual methods for standard time-to-event data. Nevertheless, this approach can lead to misguided inferences and, particularly, underestimate the standard errors of the estimated parameters. The goal is to compare these approaches with those that do not assume such imputation and that are based on the maximum likelihood estimation (MLE), the expectation-maximization (EM) algorithm, and the Turnbull estimator. All approaches will be compared using data from the Lisbon Cohort of HIV-negative of Men Who Have Sex With Men (MSM).

Keywords: Estimation of Survival, Interval Censoring, Survival Analysis.

Development of Indices to Quantify Community Capitals

Fernando Casanoves[1,2], Angie Paola Bernal Núñez[2], David Ricardo Gutiérrez Suárez[2], Héctor Eduardo Hernández Núñez[2], Isabel Gutiérrez[1], Juan Carlos Suarez Salazar[2], Raúl E. Macchiavelli[3]

[1]casanoves@catie.ac.cr, isabel.gutierrez@catie.ac.cr; CATIE-Centro Agronómico Tropical de Investigación y Enseñanza, Turrialba, Costa Rica

[2]a.bernal@udla.edu.co, da.gutierrez@udla.edu.co, h.hernandez@udla.edu.co, ju.suarez@udla.edu.co, Universidad de la Amazonia, Programa de Ingeniería Agroecológica, Facultad de Ingeniería, Florencia 180001, Colombia

[3]raul.macchiavelli@upr.edu, Colegio de Ciencias Agrícolas, Universidad de Puerto Rico, Box 9000, Mayagüez, Puerto Rico 00681

The Community Capital Framework (CCF) is a tool to improve the understanding of the resources that intervene in the development of communities. All communities have assets, resources or capitals that allow them to manage their livelihoods. These capitals are classified as: human, social, financial, built, natural, cultural, and political, and allow for a systemic analysis of community development. The CCF approach is being increasingly used in rural socio-ecological studies on natural resource management, life strategies and well-being of rural households, community tourism, technology adoption, organic certification, social vulnerability, food security, rural-ecotourism, regeneration, level organizational and community resilience, and climate change adaptation strategies.

The quantification of each capital is done with a series of indicators that vary depending on the objective of the study and the capital. The generally used methodology consists of rescaling the indicators of each capital to an interval [0,1], where 0 is the lowest value of the indicator and 1 is the highest value in the dataset. When the indicator is negative, for example, number of days without access to the farm (built capital), these are inverted by subtracting the [0,1]-rescaled indicator from 1.Because the capitals generally have different numbers of indicators, it may be necessary to rescale the sum of the indicators of each capital back to [0,1] to obtain an index score for each capital. The rescaling of these scores for each capital is intended to give each capital the same weight. Finally, the capital indices are averaged or summed to obtain a global indicator called the total capital index or welfare index.

The objective of this research is to develop a standard methodology to calculate CCF indices that considers the presence of inherently missing data, can be compared across different regions or occasions, and has a meaningful interpretation in the context of the study. Different alternatives are evaluated to address the problems of appropriate rescaling, weighting, and handling missing data. An improved method is proposed and compared to the generally used methodology to construct these indices. Possible extensions to construct other ecological indices using water and soil quality parameters is discussed.

Keywords: social studies, human wellbeing, weighted indices.

Analyzing unreplicated trials in precision agriculture

Córdoba M.[1], *Paccioretti P.*[2], *Balzarini M*[3].

[1]mariano.cordoba@unc.edu.ar, Unidad de Fitopatología y Modelización Agrícola (UFyMA), INTA – CONICET. Universidad Nacional de Córdoba, Facultad de Ciencias Agropecuarias, Departamento de Desarrollo Rural, Cátedra de Estadística y Biometría. Córdoba, Argentina.

[2]pablopaccioretti@agro.unc.edu.ar, Comisión Nacional de Actividades espaciales. Universidad Nacional de Córdoba. Instituto Mario Gulich. Facultad de Ciencias Agropecuarias, Departamento de Desarrollo Rural, Cátedra de Estadística y Biometría. CONICET. Córdoba, Argentina.

[3]monica.balzarini@unc.edu.ar, Unidad de Fitopatología y Modelización Agrícola (UFyMA), INTA – CONICET. Universidad Nacional de Córdoba, Facultad de Ciencias Agropecuarias, Departamento de Desarrollo Rural, Cátedra de Estadística y Biometría. Córdoba, Argentina.

Unreplicated trials have an important role in precision agriculture (PA). Nowadays, precise machinery promotes on-farm experimentation (OFE) to assess the effect of changes in input application rates on yields and profits in specific fields. However, limitations in OFE management may preclude more than one replicate of each treatment. Moreover, more input rates can be assessed in unreplicated trials, making OFE more efficiently carried out across fields. Thus, the single-strip treatment trial, where a field is split into two (treatment vs. control), is common in OFE. Factors such as the requirement of field-specific inference and the large number of repeated-measurement points within treatment and control plots motivate the use of permutation tests to compare treatments. However, the data in OFE are spatially correlated, and consequently, sampling is challenging. This work proposes a methodology for local control in unreplicated OFE with spatial data. A protocol integrating spatial analysis and permutation tests is developed for comparing two treatments in unreplicated trials with a probabilistic base. The methodology involves the calculation of the effective sample size (ESS) from a spatial grid defined over the trial residuals, followed by a two-hundred-cell grid random resampling of ESS and one-way permutation ANOVA to obtain the p-value comparing treatments in the randomly sampled data. The median of the empirical distribution of p-values is regarded as the expected p-value associated with the non-treatment effect hypothesis. Four example OFE datasets are used to validate the protocol. The proposed method allows measurement of the statistical significance of treated and control mean differences in unreplicated OFE. It can be extended to compare several treatments using a method for adjusting the derived p-values.

Keywords: permutation, spatial correlation, effective sample size.

Study of the performance of data-poor fish stock assessment methods

Marta Cousido-Rocha[1], *Helena Nina del Río*[2], *Santiago Cerviño*[1],*Anxo Paz*[1], *David José Nachón*[1], *Maria Grazia Pennino*[3]

[1]marta.cousido@ieo.csic.es,[1]santiago.cervino@ieo.csic.es,
[1]anxo.paz@ieo.csic.es,[1]david.nachon@ieo.csic.es, Instituto Español de Oceanografía (IEO CSIC). Centro Oceanográfico de Vigo. Subida a Radio Faro 50-52, 36390 Vigo.
[2]helenaninadel.rio@rai.usc.es, Máster en Técnicas Estadísticas, Universidade de Santiago de Compostela, A Coruña, Spain.
[3] grazia.pennino@ieo.csic.es,Instituto Español de Oceanografía (IEO CSIC). C. del Corazón de María, 8, 28002 Madrid.

Stock assessment models are mathematical and statistical techniques implemented to analyze and understand changes of fish populations. However, many of these models require a lot of information that is usually deficient for the vast majority of fish stocks. For this reason, in recent years there has been great interest in the use and development of new methods for data-poor stocks. This study focuses on two of the most used length-based methods: (1) Length Based Indicators (LBI), and (2) Length Based Spawning Potential Ratio (LBSPR). In particular, our aim is to design a specific case simulation study for the common Sole (*Solea solea*) for evaluating the LBI and LBSPR sensitivity to the uncertainty of the required life history parameter estimates and the assumptions made (selectivity, recruitment,...). To carry out the analysis the stock dynamics is simulated through operating models (OMs) which fit all relevant aspects of the fisheries system as well as how the data is collected. Our OMs follow the ideas in Fisher et al. (2020) considering plausible hypotheses about the biology of the stock, such as recruitment and growth processes, and aspects of the fishery (i.e., effort and selectivity). The performance of the LBI and LBSPR methods, in the simulation study, have led to a clear guide of how the results can be affected by the non-fulfillment of the model assumptions or by input data and life history parameters uncertainty.

Keywords: common sole, data-poor, population dynamics, simulations, uncertanity

1. Bibliography

[1] Fischer, Simon H, José AA De Oliveira, and Laurence T Kell. 2020. Linking the Performance of a Data-Limited Empirical Catch Rule to Life-History Traits. *ICES Journal of Marine Science* , 77 (5), 1914-26.

Temperature curves: a functional-circular view

Rosa M. Crujeiras[1], *Andrea Meilán-Vila*[2], *Mario Francisco-Fernández*[3]

[1]rosa.crujeiras@citmaga.gal, Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga), Universidade de Santiago de Compostela

[2]ameilan@est-econ.uc3m.es, Department of Statistics, Carlos III University of Madrid

[3]mario.francisco@udc.es, Department of Mathematics, Universidade da Coruña

Climate change, although a global problem, is usually perceived by individuals on a local scale, being changes in temperature patterns the most direct indicator of global warning. Specifically, for an Atlantic climate location (as it happens in Galicia, NW-Spain), spring and fall are expected to present mild transitions between summer/winter, but people in this region usually perceive that this transition has disappeared in the last decade. We have considered daily temperature curves along time, being each curve attached to a certain calendar day, and therefore enabling the formulation of a regression model with functional covariate (the temperature curve) and circular response (the calendar day). Such a model is fitted with observations from the early 21st century. For observations from recent years (for which the corresponding calendar day is known), the fitted model is used to predict the calendar day, observing a remarkable misalignment between observed and predicted days for a certain curve. Apart from the interesting practical results, the methodological proposal is studied, and the asymptotic bias and variance of a Nadaraya-Watson type estimator, jointly with its asymptotic distribution, is derived. Simulation results support the finite sample performance of the proposal.

Keywords: Circular data, Flexible regression, Temperature curves.

On the modelization of count data with excess zeros. Negative Binomial and Poisson regression models for COVID-19 data.

Irene de León[1], *Josu Najera-Zuloaga*[2], *Irantzu Barrio*[3], *Jose María Quintana*[4]

[1]irene.deleonperla@gmail.com, Department of Mathematics University of the Basque Country UPV/EHU

[2]josu.najera@ehu.eus, Department of Mathematics University of the Basque Country UPV/EHU

[3] irantzu.barrio@ehu.eus, Department of Mathematics University of the Basque Country UPV/EHU; Basque Center for Applied Mathematics, BCAM

[4] josemaria.quintanalopez@osakidetza.eus, Osakidetza Basque Health Service, Galdakao-Usansolo University Hospital; Network for Research on Chronicity, Primary Care, and Health Promotion (RICAPPS)

Count data are common in general and in clinical practice in particular. In addition, it is common to find count variables with a high presence of zeros, known as zero inflated distribution.

Count outcomes are often modeled using the Poisson distribution. However, the Poisson distribution assumes that mean and variance are equal, however, this is not always the case in practice, resulting that the variance is significantly greater/smaller than the mean (known as overdispersion/underdispersion). In this case, the Negative Binomial distribution can be used instead. Furthermore, it is common in count data to encounter an apparent excess of zeros, often with respect to the Poisson and even with respect to the Negative Binomial distribution. In this context, zero-inflated regression models provide a better modeling of the data.

The goal of this work is to study and compare the performance of Poisson and Negative Binomial regression models (considering or not zero inflated models). We have applied this model to a real data set of a total of 380,074 adult patients infected with the virus SARS-CoV-2 from March 1, 2020 to January 9, 2022 in the Basque Country. We have analyzed two discrete outcome variables: a) number of reinfections that each individual has had during that period, and b) number of days admitted to the hospital after the first infection.

Keywords: Poisson regression, Negative Binomial regression, Zero-inflated models

Air quality analysis with supervised learning algorithms in Coyhaique, Chile

Mailiu Díaz Peña[1], *Felipe A. Medina Marín*[2], *Ana Karina Maldonado Alcaíno*[3], *Dante Cáceres Lillo*[4]

[1]mailiu.diaz@unab.cl, Facultad de Ingeniería, Universidad Andrés Bello
[2]f.medina@uchile.cl, Instituto de Salud Poblacional, Facultad de Medicina, Universidad de Chile
[3]vgtariana@uchile.cl, Instituto de Salud Poblacional, Facultad de Medicina, Universidad de Chile
[4]dcaceres@uchile.cl, Instituto de Salud Poblacional, Facultad de Medicina, Universidad de Chile

Coyhaique is one of the most polluted cities in Latin America, located in the southern region of Chile, with a population density of no more than 7 inhabitants per km^2. Currently, it is among the three cities with the highest concentrations of particulate matter 2.5 (PM2.5) in the country, mainly due to the use of firewood as a source of energy and heating, negatively impacting the quality of life of those who reside in the area. However, meteorological factors also influence the levels of PM2.5 concentration, so it is relevant to quantify their importance. To do this, we implemented supervised learning algorithms using the data available in the last 10 years from fixed air quality monitoring stations in Coyhaique. We used these models to predict the favorable conditions under which PM2.5 exceeds the Chilean standards, which define a limit of 20 $\mu g/m^3$. Cross-validation was used with different performance metrics to quantify the accuracy and precision of the fitted models. The evaluation allows for explaining the trend of PM2.5 concentrations and predicting its behavior, which can help decision-makers optimize resource allocation to comply with air quality standards.

Keywords: particulate matter, machine learning.

Modelling the COVID-19 ICU occupancy with area-level random regression coefficient Poisson models

Naomi Diz-Rosales[1], *María José Lombardía*[2], *Domingo Morales*[3]

[1]naomi.diz.rosales@udc.es, CITIC, Universidade da Coruña, Spain.
[2]maria.jose.lombardia@udc.es, CITIC, Universidade da Coruña, Spain.
[3]d.morales@umh.es, IUCIO, Universidad Miguel Hernández de Elche, Spain.

COVID-19 has had disastrous consequences in all key areas of human welfare. The information and tools currently available to fight the virus, with vaccination as the main banner, have practically normalised the disease. However, the evolution of different variants, together with the weakening of the health system after years of continuous pressure and the predicted emergence of new epidemics in the not too distant future, highlight the need to develop explanatory and predictive tools that allow temporal and spatial monitoring of the rate of infection and the resulting overload of care.

Faced with the challenge posed by the scarcity and lack of homogeneity of the data collected, especially during the first year of the pandemic, the prevailing unknown nature of the virus, as well as the territorial heterogeneity in both the spread and severity of the disease, the mixed models in Small Area Estimation (SAE) are shown to be a methodological approach with great potential. These techniques can make a significant contribution to health planning, resource allocation and the implementation of non-pharmacological intervention measures.

In this context, we propose the development of random slope mixed models with the target of modelling the ICU (Intensive Care Unit) occupancy rate due to COVID-19. Given that the collapse of ICUs constitutes one of the main bottlenecks in care capacity, the estimation of this proportion is presented as one of the indicators with the greatest explanatory capacity of the overload experienced in different scenarios of contagion and severity of the disease.

Specifically, under an area-level random regression coefficient Poisson model, this work derives area-level estimators of occupied ICU bed counts and occupancy ratios, introducing bootstrap estimators of the mean squared errors. The maximum likelihood estimators of the model parameters and the mode predictors of the random effects are calculated by a Laplace approximation algorithm. Simulation experiments are implemented to investigate the behavior of the fitting algorithm, the predictors and the mean squared error estimator. The new statistical methodology is applied to aggregated data at the level of the 11 Health Areas of the Spanish Autonomous Community of Castilla y León, corresponding to a time range bounded between 2020 and 2022, defining the domains as the health area-day crossover. In these spatio-temporal areas, the explanatory capacity of the model is evaluated and an initial analysis of its predictive value is provided.

Keywords: COVID-19 ICU occupancy, Small Area Estimation, random coefficient Poisson regression models

Smoothing estimation of the functional ROC Curve

Graciela Estévez-Pérez[1]

[1]graciela.estevez.perez@udc.es, Departamento de Matemáticas, Universidade da Coruña

Technical development over the last few decades has resulted in the emergence of complex data, in many cases functional data (FD). This type of data can emerge in many medical studies which are geared towards detecting diseases, predicting their course or evaluating the response to a therapy, to name a few. Thus, it is very useful to have statistical methods enabling us to evaluate diagnostic tests based on functional biomarkers. Estévez and Vieu (2021) developed recently a diagnostic test that use functional variables as biomarkers and proposed an empirical estimate of the functional ROC curve. In order to improve this methodology, the present paper proposes a procedure to obtain a smooth version of nonparametric estimator of ROC curve.

In addition, a comprehensive simulation study was carried out to investigate the discriminatory and predictive abilities of the new functional diagnostic tests and the effect of the bandwidth parameter over them. Special attention was paid to the improvement achieved using the kernel estimators versus the empirical one. The empirical study was undertaken to examine the impact of estimation method of ROC curve in the functional context and their interaction with the other parameters involved in diagnostic test: type of representative curves and semi-metric for the projection. This effect is measured by on the discriminatory power of the test, through numerical indexes like the AUC, and on its predictive ability through misclassification rates.

Finally, to illustrate the methodology of functional diagnostic test and to complete the comparison among kernel estimation methods and the empirical ROC curve, the analysis of two real medical data sets has been carried out: one deals with gene expression levels for tumoral/normal samples of prostate cancer (data published by Singh et al. (2002) and included in the `R` package `depthTools`); the other dataset is about white matter structures in the brain in multiple sclerosis patients (dataset included in the `R` package `refund`, (Goldsmith et al., 2016)).

Keywords: Functional biomarker; ROC curve; Kernel smoothing.

REFERENCES

Estévez-Pérez, G. & Vieu, P. (2021). A new way for ranking functional data with applications in diagnostic test. Computational Statistics, 36(1), 127-154.

Goldsmith, J., Huang, L., Scheipl, F., Gellar, J., Harezlak, J., McLean, M. W., ... & Reiss, P. (2016). refund: Regression with Functional Data. R package version 0.1-14

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., ... & Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer cell, 1(2), 203-209.

Software tools at the Biostatnet-Granada node

P. Femia[1,2], MA Montero[1,2]; J. Melchor[1,2], MA Luque[1], JD. Luna1[1,2], J. Martorell[1,3], JA. Villatoro[1,3], P. Carmona-Saez[1,3]

[1]Dpt. Statistics and Operations Research. University of Granada, Spain
[2]Instituto de Investigacion Biosanitaria, ibs.GRANADA, Granada, Spain
[3]Centre for Genomics and Oncological Research (GENYO) Pfizer, Granada, Spain

The Biostatnet node in Granada integrates researchers from six different institutions and universities. They are actively working to develop new statistical and machine-learning methods for analyzing biomedical data. An important aspect of their research lines is the deployment of software applications that make these methods available to the scientific community. Here we provide an overview of the main software applications available from this research group. These include dedicated databases, web-based applications, and R packages, covering a broad range of research needs.

Bioinformatics software tools

- ImaGEO: A web tool that implements a complete and comprehensive meta-analysis workflow to analyze gene expression datasets from GEO identifiers. [1]
- MetaGENyO: A web tool that implements a complete workflow for meta-analysis in Genetic Association Studies. [2]
- DExMA: R Package for Differential Expression Meta-Analysis. [3]
- mCSEA. R package that implements a GSEA-based approach to detect differentially methylated regions and gene expression-methylation data integration [4]

Agreement among raters

- Delta: a stand-alone application that implements the model *Delta* for evaluating the agreement between two raters on a nominal scale and the model *delta* for assessing the level of knowledge in a multiple-choice test [5]

Teaching Software

- BioestadísticaR: An R package for basic biostatistics teaching in medical sciences [6]

Keywords: Statistical Software, Web-based applications, R-packages.

REFERENCES

1. Toro-Domínguez, D., Martorell-Marugán, J., López-Domínguez, R., García-Moreno, A., González-Rumayor, V., Alarcón-Riquelme, M.E., and Carmona-Sáez, P. (2019). ImaGEO: integrative gene expression meta-analysis from GEO database. Bioinformatics *35*, 880–882. 10.1093/bioinformatics/bty721.
2. Martorell-Marugan, J., Toro-Dominguez, D., Alarcon-Riquelme, M.E., and Carmona-Saez, P. (2017). MetaGenyo: a web tool for meta-analysis of genetic association studies. BMC Bioinformatics *18*, 563. 10.1186/s12859-017-1990-4.
3. Villatoro-García, J.A., Martorell-Marugán, J., Toro-Domínguez, D., Román-Montoya, Y., Femia, P., and Carmona-Sáez, P. (2022). DExMA: An R Package for Performing Gene Expression Meta-Analysis with Missing Genes. Mathematics *10*, 3376. 10.3390/math10183376.
4. Martorell-Marugán, J., González-Rumayor, V., and Carmona-Sáez, P. (2019). mCSEA: detecting subtle differentially methylated regions. Bioinformatics *35*, 3257–3262. 10.1093/bioinformatics/btz096.
5. Femia, P. et al. (2022). Delta. Delta. https://www.ugr.es/~bioest/software/cmd.php?seccion=agreement.
6. Femia Marzo, P.J., Carmona Sáez, P., Luna Del Castillo, J.D.D., Melchor Rodríguez, J.M., Acal González, C.J., Romero Béjar, J.L., Expósito Ruiz, M., Villatoro García, J.A., and Montero Alonso, M.Á. (2022). BioestadísticaR versión 2.0. https://digibug.ugr.es/handle/10481/77064

Impact of imbalance data on Logistic Regression models to predict risk plant disease

Fiore Juan Manuel[1], *Suarez Francor*[2], *Balzarini Monica*,[3] Bruno Cecilia[4]

[1]juanmfiore@mi.unc.edu.ar, Unidad de Fitopatología y Modelización Agrícola (UFyMA). INTA-CONICET

[2]suarezfranco@agro.unc.edu.ar, UFyMA. INTA-CONICET

[3]monica.balzarini@unc.edu.ar, UFyMA. INTA-CONICET. Estadística y Biometría. Facultas de Ciencias Agropecuarias (FCA). Universidad Nacional de Córdoba (UNC). Argentina.

[4]cebruno@agro.unc.edu.ar, UFyMA. INTA-CONICET. Estadística y Biometría. FCA. UNC. Argentina.

The imbalance problem occurs when there is a skewed class distribution, large samples of one class, and few samples of the other class. The degree of imbalance can vary largely and when this happens the ability of the statistical model to predict the occurrence of the minority class is heavily affected. This situation is frequent in plant diseases, with the number of unhealthy plants being lower than that of healthy plants. Logistic regression (LR) models are linear models that allow us to predict a binary event from multiple predictive variables. This work aims to evaluate the impact of unbalance data on two different models: a classical LR and a more state-of-the-art version of a Logistic regression in this case boosted (BLR), on a binomial classification problem from plant disease data using climatic variables as predictors. The current study implements two under-sampling algorithms, a well-known Tomek algorithm and Condensed K-neighbours (CNN), and two oversampling algorithms: Synthetic minority oversample Technique (SMOTE) and Adaptative Synthetic Sapling (ADASYN) with three different proportions of imbalance based on two pathosystems collected on cucurbits crops. The original imbalance dataset was 5:1 in the PRSV pathosystem and 9:1 in the CMV, absent to presence, respectively. Then, we generated different levels of imbalance to change the relationship between the absence/presence of pathosystem. Thus, for each original dataset we modify the dataset with 10:3 ratio, other with 10:6 ratio, and finally 1:1 ratio (completely balanced). The original dataset was partitioned into 80% for training and 20% for validation. Then, the two models, LR and BLR, were trained and cross-validated. The experiment was reproduced six times with different seeds to ensure the reproducibility and randomness on the dataset partition for validations. The criteria to compare the model's performance to predict the risk of plant disease were the Area Under the Curve (AUC) from the receiver operator characteristic (ROC) curve, sensitivity and specificity were used as metrics. Finally, the models were tested with the remaining 20% dataset who did not participate in the training. LR with the original (imbalance) dataset was the worst performance with an AUC of 0.62, and a sensitivity of 0.28 although its specificity achieved about 0.95. LBR had an AUC of 0.75, a sensibility of 0.53, and a Specificity of 0.98. The model that most responded to the different imbalance levels was the LR. Both algorithms got the best results with SMOTE 6:10. LR AUC improved to 0.76 and LBR achieved and AUC of 0.82. leaving the 1:1 ratio in second place. This pattern of SMOTE 6:10 performing better, repeats itself in combination with the under-sampling techniques. Indicating there is an optimum threshold in which from that point the excess data produced by the algorithm generates noise. Undermining the algorithms performance.

Keywords: oversampling algorithms, under-sampling algorithms, cucurbits crops.

Correction for baseline covariates in clinical trials and observational studies

Matilde Francisco[1], *Klaus Langohr*[2]

[1]fc57478@alunos.fc.ul.pt, Faculty of Sciences, University of Lisbon
[2]klaus.langohr@upc.edu, Department of Statistics and Operations Research, Polytechnic University of Catalonia

Longitudinal studies allow the repeated monitoring of health outcomes or risk factors, and the identification of differences in outcomes. Baseline measurements, demographic characteristics or measurements taken at the beginning of the study of the response variable or variables correlated with it, have several purposes, including the assessment of treatment effect based on the change from baseline. The question on whether to consider baseline as a covariate or a dependent variable is frequently asked.

Not accounting for baseline, can not only affect the magnitude of differences detected, but also the direction of these differences, which can result in different clinical conclusions. However, correcting for baseline effects can also introduce bias, which is problematic, especially in cases where sample sizes are small. The lack of consistency in the literature around this topic contributes to the difficulty to establish a standard statistical approach, so studies' specific characteristics influence the decision on what statistical approach should be used.

In clinical trials, randomization is a fundamental process, since it helps prevent bias associated with the selection of candidates receiving each treatment. It ensures that it is possible to compare the effect of the different treatments between groups, since they are similar in almost every other critical aspect. In observational studies, the lack of randomization can result in selection bias, since the baseline characteristics of individuals in exposure groups may differ. If these characteristics have a significant role in predicting the outcome, their imbalance between groups can cause bias.

When adjusting a model, it is necessary to adopt different strategies regarding the use of baseline values. The use of Analysis of covariance (ANCOVA) has been advocated when there is the need to correct for baseline. In this model, under the assumption of the existence of a correlation between baseline and post-baseline measurements, the baseline values are included in the model as a covariate, and the post-baseline values are the response variable. The constrained longitudinal data analysis (cLDA) is built under the assumption that the randomization of the subjects involved was efficient, so it is assumed that means at baseline are identical for the groups being compared. Due to this, both baseline and post-randomization values are dependent variables.

In this work, the models previously described are going to be applied to a data set from the Safety and efficacy of cognitive training plus epigallocatechin-3-gallate in young adults with Down's syndrome (TESDAD) clinical trial, in order to access if different approaches to baseline have impact on the findings.

Keywords: Baseline-value adjustment, linear mixed models, longitudinal studies.

Mapping MusiQoL onto the EQ-5D-5L utility index in patients with multiple sclerosis

Leire Garmendia Bergés[1,2,3], *Iñigo Gorostiza Hormaetxe*[1,2,3], *Alfredo Rodríguez Antigüedad*[4,5], *Mar Mendibe Bilbao*[4,5], *Irantzu Barrio Beraza*[2,6], *Amaia Bilbao Gonzalez*[1,2,3]

[1]lgarmendia@kronikgune.org, Kronikgune Institute for Health Services Research, Barakaldo, Spain

[2]Network for Research on Chronicity, Primary Care, and Health Promotion (RICAPPS), Bilbao, Spain

[3]Osakidetza Basque Health Service, Basurto University Hospital, Research and Innovation Unit, Bilbao, Spain

[4]Osakidetza Basque Health Service, Cruces University Hospital, Department of Neurology, Barakaldo, Spain

[5]BioCruces Bizkaia Health Research Institute, Barakaldo, Spain

[6]Department of Mathematics, University of the Basque Country, Leioa, Spain

Multiple sclerosis is a chronic disease that has a high impact in the quality of life. Due to the prevalence and the economic cost of the disease, the cost-effectiveness analyses are very relevant to decision makers. A measure that is commonly used in those studies are the quality-adjusted life-years (QALYs), which combine quantity and quality of life and allow comparing the health related quality of life (HRQoL) of different treatments, or populations. One of the necessary parameters to estimate QALYs is the utility for a given health state, as a measure of quality of life. Utilities are usually measured with generic health-related quality of life questionnaires, such as the EQ-5D-5L questionnaire, from which a utility index is derived based on a preference-based scoring function. However, in clinical practice, specific questionnaires, such as the multiple sclerosis international quality of life (MusiQoL), are more often used instead of the generic ones. Unfortunately, it is not possible to obtain the utility index directly from the specific questionnaires, and for that, mapping functions need to be developed to relate specific and generic questionnaires, in order to obtaining the utility index based on the specific questionnaires.

The objective of this study is to develop mapping functions to map the global score of the MusiQoL questionnaire onto the EQ-5D-5L utility index in patients with multiple sclerosis, using different algorithms and to compare and validate those functions.

To develop these mapping functions we used a real data set of patients with multiple sclerosis. We recruited 185 patients and collected baseline information using the EQ-5D-5L and the MusiQoL questionnaires, together with some sociodemographic questions. 6 months later, only 165 of the patients completed the follow-up questionnaires. For the development of these mapping functions baseline data were used, and for the validation the 6-month follow-up data. To map the global score of the MusiQoL into the EQ-5D-5L utility index, general linear models (GLMs), Tobit models, Beta regression models, and adjusted limited dependent variable mixture models (ALDVMM) were used.

This study provides different mapping algorithms to predict the EQ-5D-5L utility index based on the global score of the MusiQoL questionnaire in patients with multiple sclerosis.

Keywords: Mapping, utility index, multiple sclerosis.

Estimating the probability of discharge among Covid-19 hospitalizations using cure models

Ida Höglund Persson[1,2], *Klaus Langohr*[3], *Guadalupe Gómez Melis*[3]

[1]idaper@chalmers.se, Department of Statistics and Operations Research, Universitat Politécnica de Catalunya

[2] Department of Mathematical Sciences, Chalmers University of Technology

[3]klaus.langohr@upc.edu, lupe.gomez@upc.edu, Department of Statistics and Operations Research, Universitat Politécnica de Catalunya

This Master thesis is focused on cure models and their applicability in handling survival models with long-term survival due to immunity. These models have gained increased attention in recent years due to medical advances within several disease treatments. Cure models are indicated whenever a dataset contains a large proportion of right-censored individuals at the follow-up time that can be suspected to include individuals who will never experience the event of interest due to being cured (or immune).

Cure models assume that the population is divided into two populations - one that is susceptible to the disease, and one that is immune to the event. In this study, the main objective is to estimate the probability of cure, referred to as the incidence. Various cure models will be investigated and applied to a dataset comprising 4000 hospitalized Covid-19 patients from the metropolitan south area of Barcelona during four waves of the pandemic. The purpose of the study is to explore which cure models are more appropriate to model the behaviour of Covid-19 patients concerning time to death and time to discharge

Mixture cure models allow the subpopulations to have different survival distributions while non-mixture cure models are easily interpreted due to similarities with proportional hazards models. Estimation of the survival of the uncured population, called the latency function, has mainly been done parametrically. However, recent research provides new complex and accurate non-parametric methods that will be applied in this study. A comparison of the various estimation methods is provided with a discussion of the advantages and disadvantages of each one of the approaches.

Due to the recent increase in the number of applications within cure models, an extension has been suggested considering a background survival for the overall mortality of the non-sick population. Previous research within cure models has mainly examined children as it is a group with a low background mortality rate or a nonfatal outcome such as disease recurrence. Since this dataset consists of adult Covid-19 patients this extension is an option to contemplate.

Keywords: Cure models, cure proportion, Covid-19

High-dimensional Unsupervised and/or Supervised problems: a distance-based depth prototypes fuzzy approach

Itziar Irigoien[1], *Susana Ferreiro*[2], *Basilio Sierra*[3], *Concepción Arenas*[4]

[1]itziar.irigoien@ehu.es, Department of Computation and Artificial Intelligence, University of the Basque Country (UPV/EHU)

[2]susana.ferreiro@tekniker.es, Intelligent Information Systems Unit, Tekniker

[3]b.sierra@ehu.eus, Department of Computation and Artificial Intelligence, University of the Basque Country (UPV/EHU)

[4]carenas@ub.edu, Statistics Section of the Department of Genetics, Microbiology and Statistics, University of Barcelona (UB)

Supervised and unsupervised classifications are crucial in many areas such as biomedicine, or industry, among others, where different types of data sets and high-dimensional data (number of features p larger than the number of observations n, $p >> n$) are common. A very important issue is that some units are more typical of the group they belong to than others. Following the ideas of Hastari et al. (2020) We propose a new fuzzy supervised classification approach based on the construction of prototypes from an objective function that incorporates the information from the class labels, as well as a distance-based depth function to fuzzify the partition. As it is a fuzzy methodology, the objective function contains a term related to the entropy of the memberships. Obtaining the prototypes, label prototypes, and weighted memberships relative to each class is carried out by an iterative scheme. The method has hyperparameters that need to be tuned and a grid search is used. Then, when the approach is supervised, the selection of the hyperparameter is made according to an adequate metric (accuracy rate, for instance) reached on a k-fold cross-validation setting. When the approach is non-supervised, there is a lack of an external validation variable and, as an internal validation approach, a permutation approach related to the Gap statistics leads to good results. The procedure is distance-based, so it can be used in data sets of a very varied nature, particularly without restrictions with high-dimensional data and with data sets where the Euclidean distance is not suitable, but other distances are. Notably, it may be a very good choice for functional data. Furthermore, the method selects the prototypes among the deepest units preventing the prototype from not belonging to the sample, as can happen in the case of centroids. Its performance on synthetic data sets along with real data showed good rates of correct classification and it is competitive with other methods. The proposed method provides an interesting alternative to other fuzzy clustering purposes.

Keywords: Fuzzy classification, Prototypes, Depth function.

P. Ashtari, F. N. Haredasht, H. Beigy (2020). Supervised fuzzy partitioning, Pattern Recognition, 97, Article 107013.

Dealing with spatiotemporal dependence in spatial stock assessment models

Francisco Izquierdo[1], *Marta Cousido-Rocha* [1], *Santiago Cerviño*[1] *& Maria Grazia Pennino*[2]

[1]francisco.izqtar@gmail.com, Department of Fisheries, Instituto Español de Oceanografía (IEO-CSIC, C.O. de Vigo)

[2] Instituto Español de Oceanografía (IEO-CSIC), Centro Oceanográfico de Madrid, C. del Corazón de María, 8, 28002 Madrid, Spain.

Assessing the state of commercial fish stocks is a complex task that requires understanding how species abundance changes over time and space. Spatial stock assessment models can account for the large-scale spatiotemporal dependence of ecological and fishing processes. However, they often fail to consider small-scale spatial correlation unless this is addressed in their input information, such as relative abundance indices. Indices of relative abundance are one of the most important sources of information for stock assessment models, as they are used to calibrate population trends. For many stocks, fishery-independent survey data are not available, so a common practice is to derive abundance indices from fishery-dependent catch per unit effort (CPUE). CPUE indices are known to be influenced by several factors, among which spatiotemporal factors are very relevant, as they affect processes such as species reproduction or feeding patterns. In this study, we used the integrated nested Laplace approximation (INLA) to fit three different CPUE models to simulated lattice data of yellowfin tuna (*Thunnus albacares*) in the Indian Ocean: 1) a model in which spatial dependence was included as a random effect, 2) a model with a spatial Besag component and 3) a model with a spatial Besag component interacting with time through an autoregressive process of order 1. The best CPUE index selected was finally used to calibrate the Stock Synthesis spatial assessment model for the yellowfin tuna stock.

Keywords: CPUE, correlation, INLA, population models, yellowfin

Modelling spatial distribution of small pelagic fishes' biomass using boosted regression trees hurdle models

Laura Julià[1], *Maria Grazia Pennino*[2], *Jose M. Bellido*[3], *Marta Coll*[4,5] *and Francisco Ramírez*[5]

[1]lauraj@icm.csic.es, Department of Renewable Marine Resources, Institute of Marine Sciences (ICM-CSIC), Barcelona, Spain

[2]Instituto Español de Oceanografía (IEO-CSIC), Centro Oceanográfico de Vigo, Spain

[3]Instituo Español de Oceanografía (IEO-CSIC), Centro Oceanográfico de Murcia, Spain

[4]Ecopath International Initiative Research Association, Barcelona, Spain

[5]Department of Renewable Marine Resources, Institute of Marine Sciences (ICM-CSIC), Barcelona, Spain

Small pelagic fishes (SPFs) in the northwestern Mediterranean are key elements of marine ecosystems: they represent a significant fraction of fish catches and are the main source of food for a large number of species. Evaluating the spatial distribution of the biomass available for these species allows to assess the food availability and enables to make conservation decisions. In this study, we implemented species distribution models (SDMs) for the most abundant SPFs, the European anchovy (*Engraulis encrasicholus*), the European sardine (*Sardina pilchardus*) and the sardinella (*Sardinella aurita*), in order to predict spatially-explicit biomasses. The study area includes the waters of the Iberian continental shelf within the geographical sub-area (GSA) 06. Species data comprises surveys carried out between June and July during the period 1998 to 2017 (both included) by the MEDITS (MEDIterranean Trawl Survey) oceanographic surveys. The SDMs were adjusted using Boosted Regression Trees (BRT) technique, defining sea surface temperature (ºC), sea surface salinity (PSU), net primary productivity (mg/m3), sea depth (meters) and year as predictor variables of the models. To deal with spatial autocorrelation, an autocovariable derived from the residuals of the model, was included to the environmental variables in the model (BRT-RAC method). Since the biomass had a long-tailed and zero-inflated distribution, we also used hurdle models. This approach deals with high numbers of zeros by using a two-step modelling process. In the first one, presence/absence data were modelled using a Binomial distribution in order to obtain a prediction of presence probability of the studied species; in the second, biomass data were modelled using a Gaussian distribution only in areas where species were expected to be present. Finally, we predicted the annual (mean of June and July) probability of occurrences as well as the back-transformed biomass for the years between 1998 and 2019 and then we combined both of them. As a result, our study determined the environmental influence on the biomass distribution of the three species, identified important marine areas for each species, and evaluated their species distribution trends over last decades.

Keywords: Species distribution modelling, BRT-RAC hurdle models, small pelagic fishes

COVID-19 reinfections study from different statistical approaches

Nere Larrea[1,2,3], *Leire Garmendia*[2,3,4], *Irantzu Barrio*[3,5,6], *Klaus Langohr*[7], *Cristian Tebe*[8], *Guadalupe Gómez Melis*[7], *Jose María Quintana*[1,3]

[1]nlarrea@kronikgune.org, Research Unit Hospital Galdakao-Usansolo, Galdakao, Spain

[2]Kronikgune Institute for Health Services Research, Baracaldo, Spain

[3]Network for Research on Chronicity, Primary Care, and Health Promotion (RICAPPS)

[4]Research and Innovation Unit Hospital Basurto, Bilbao, Spain

[5]Department of Mathematics, University of the Basque Country UPV/EHU, Spain

[6]BCAM, Basque Center for Applied Mathematics, Vizcaya, Spain

[7]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya-BarcelonaTECH, Barcelona, Spain

[8]Institut d'Investigació Biomédica de Bellvitge (IDIBELL), L'Hospitalet, Spain

Recovery from a SARS-CoV-2 infection might confer some degree of immunity to reinfection for a period of time. However, a small percentage of patients contracts (or reinfects from) COVID-19 a second time. A total of 380,074 adult patients were infected with SARS-CoV-2 from March 1, 2020 to January 9, 2022 in the Basque Country; from those, 10,968 (2.89%) were reinfected.

The aim of this study is to analyse which factors are related to the probability of reinfection. We are as well interested in studying the effect that different transition paths after the first infection might have on a worst prognosis after reinfection, taking into account patient's sociodemographic, comorbidity and other clinically relevant factors described in the literature.

The use of different statistical techniques allow for a more comprehensive analysis of the data, providing insights into the complex relationships between the various factors. Logistic regression is used to examine the association between individual factors and the likelihood of reinfection, while the competing risks model is used to assess the time to reinfection and its impact on outcomes considering death as a competing risk. Finally, multistate models are employed to estimate different transition probabilities, such as the probability from infection to hospital admission or from infection to ICU.

Summarizing, this is a preliminar population study on reinfected SARS-CoV-2 patients showing, on one hand, those factors associated with a reinfection and, on the other, which is the likelihood of a worst prognosis among reinfected patients.

Keywords: Reinfections, multistate models, competing risks model.

Utility of the integrative analysis for the identification of microRNA for diagnosis

Julieth López-Castiblanco[1], *David Niño*[2], *Liliana López-Kleine*[3], *Adriana Rojas*[4], *Litzy Bermúdez*[5]

[1]julalopezcas@unal.edu.co, Department of Statistics, National University of Colombia

[2]jdninot@unal.edu.co, Department of Statistics, National University of Colombia

[2]llopezk@unal.edu.co, Department of Statistics, National University of Colombia

[2]rojas-adriana@javeriana.edu.co, Genetics Human Institute, Pontificia Javeriana University

[2]litzybermudez@javeriana.edu.co, Genetics Human Institute, Pontificia Javeriana University

Lung cancer is one of the main causes of mortality due to the late detection of this disease. Currently, methods to identify lung cancer are found in lung cancer tissues. Non-invasive tools for early detection would be very helpful. New studies have achieved some possibilities of diagnosis using blood samples. Nevertheless, knowledge about biomarkers in blood and circulating exosomes is still very sparce. Therefore, this study seeks to identify which miRNA present in exosomes of blood sample share deferentially expressed in people with lung cancer through biostatistical analysis to support diagnosis of lung cancer. For this aim, the work was focused on one of the most common kind of lung cancer: adenocarcinoma. Two publicly available databases of the National Center for Biotechnology Information (NCBI) were used (accession numbers: GSE71661 and GSE111803). Both databases were obtained through non-coding RNA profiling by high throughput sequencing experiment, so both have discrete data and thus, can be analyzed with methodologies adapted to this, like DESEQ2.

Initially, quality control was done to conduct a differential gene expression analysis based on the negative binomial distribution. This methodology allowed to find five common miRNA between both databases which can help to characterize lung adenocarcinoma. Consequently, the identification of genes known to be targets of these miRNAs was done using miRNet. Then, the target genes were analyzed for their functional role and highlighted metabolic pathways using DAVID. With this enrichment analysis we were able to construct a network with the five differential expressed miRNA and their gene targets. Furthermore, topological analysis of the network allowed to identify potential important genes in adenocarcinoma.

Keywords: microRNA, lung cancer.

Statistical models for forensic voice comparison from a phonetic acoustic approach.

Fernanda López-Escobedo [1], *N. Sofía Huerta-Pacheco*[2]

[1]flopeze@unam.mx, Escuela Nacional de Ciencias Forenses, Universidad Nacional Autónoma de México

[2]nshuerta@enacif.unam.mx, CONACYT - National School of Forensic Sciences, Universidad Nacional Autónoma de México

In this work, two models developed by Rose et al. (2004) and Morrison (2011) were implemented for acoustic-phonetic data from speech recordings. These methods have been commonly used to assess the strength of evidence from an auditory-acoustic-phonetic approach. The results of both models are numerical values that can be used to evaluate the evidence's strength in the likelihood ratio framework.

In many real-world cases, the amount of voice data is not sufficient to use models commonly used in automatic speaker recognition, such as Gaussian mixture models (GMMs) or Deep Neural Network (DNN). For this reason, we chose to test the Rose et al. (2004) and Morrison (2011) models and explore their performance when only a small number of samples are available. The model proposed by Rose et al. (2004) represents the distribution of the variables with normal curves because, after studying the behavior of the acoustic parameters in a corpus of 60 Japanese speakers, they were found not to have a distribution sufficiently far from normal to warrant non-parametric modeling. In contrast, the model proposed by Morrison (2011) assumes a distribution of variables with normal curves when measured across samples of the same speaker, but unlike Rose et al. (2004), it assumes a nonparametric distribution of variables when measured across different speakers.

In this work, a speech corpus was collected from 27 female speakers with an average recording time of approximately 2:13 minutes. Each vowel was manually segmented, and different metrics of the first four formants (F1, F2, F3, and F4) of the five Spanish vowels /a/, /e/, /i/, /o/, and /u/ were analyzed. After a descriptive analysis of the data and due to the high variability, it was decided to remove outliers with a filter based on 95% interval confidence. The likelihood ratio values obtained in the case of the same speaker pairs of recordings are in agreement in both models. This shows that the results of both models are consistent and can be used when the number of speech samples is limited.

Keywords: Forensic voice comparison, Likelihood ratio, Acoustic-phonetic data

Bibliography

Rose, P., Lucy, D., & Osanai, T. (2004). Linguistic-Acoustic Forensic Speaker Identification with Likelihood Ratios from a Multivariate Hierarchical Random Effects Model-A Non-Idiot's Bayes' Approach. Proceedings of the 10th Australian International Conference on Speech Science and Technology.

Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM). Speech Communication, 53(2), 242-256.

Colliders in Applied Biomedical Research: an educational interactive web application

Miguel Ángel Luque-Fernandez[1,2], Pedro Femia[1], Miguel Ángel Montero-Alonso[1], Juan M. Melchor[1], Pedro Carmona-Saez[1], Juan de Dios Luna del Castillo[1].

[1]Department of Statistics and Operations Research, University of Granada, Granada, Spain
[2]Department of Non-Communicable Disease Epidemiology. London School of Hygiene and Tropical Medicine. London, U.K.

Classical biomedical research (Epidemiology / Biostatistics) has focused on the control of confounding, but it is only recently that applied researchers have started to focus on the bias produced by colliders. A collider for a certain pair of variables (e.g., an outcome Y and an exposure A) is a third variable (C) that is caused by both. In DAGs terminology, a collider is the variable in the middle of an inverted fork (i.e., the variable C in A → C ← Y). Controlling for, or conditioning an analysis on a collider (i.e., through stratification or regression) can introduce a spurious association between its causes. This potentially explains many paradoxical findings in the medical literature, where established risk factors for a particular outcome appear protective. We used an example from non-communicable disease epidemiology to contextualize and explain the effect of conditioning on a collider. We generated a dataset with 1,000 observations and ran Monte-Carlo simulations to estimate the effect of 24-hour dietary sodium intake on systolic blood pressure, controlling for age, which acts as a confounder, and 24-hour urinary protein excretion, which acts as a collider. We illustrate how adding a collider to a regression model introduces bias. Thus, to prevent paradoxical associations, applied researchers estimating causal effects should be wary of conditioning on colliders. We provide R-code in easy-to-read boxes throughout the manuscript and a GitHub repository (https://github.com/migariane/ColliderApp) for the reader to reproduce our example. We also provide an educational web application allowing real-time interaction to visualize the paradoxical effect of conditioning on a collider http://watzilei.com/shiny/collider/. We investigated a situation where, adding a certain type of variable to a linear regression model, called a "collider", led to bias with respect to the regression coefficient estimates while still improving the model fit. DAGs are based on subject matter knowledge and are vital for identifying colliders. Determining if a variable is a collider involves critical thinking about the true unobserved data generation process and the relationship between the variables for a given scenario. Then, the decision whether to include or exclude the variable in a regression model using observational data in biomedical research is based on whether the purpose of the study is prediction or explanation/causation. Under the structures we investigated here, adding a collider to a regression model is not advised when one is interested in the estimation of causal effects, as this may open a back-door path. However, if prediction is the purpose of the model, the inclusion of colliders in the models may be advisable if it reduces the model's prediction error. Most research in Epidemiology and Biostatistics tries to explain how the world works (i.e., it is causal), thus, to prevent paradoxical associations, applied researchers estimating causal effects should be aware of such variables.

Keywords: Colliders.

References.

Pearl, J. Causality: models, reasoning, and inference (2nd ed.). Cambridge: Cambridge University Press. 2009.

Luque-Fernandez MA, Zoega H, Valdimarsdottir U, Williams MA. Deconstructing the smoking-preeclampsia paradox through a counterfactual framework. Eur J Epidemiol. 2016; 31: 613-623.

Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2009; 39:417-420.

Response surface survival analysis methodology for block design

Oscar Orlando Melo Martínez[1], *Ana Patricia Chávez*[2], *Nelson Alirio Cruz* [3]

[1]oomelom@unal.edu.co, Department of Statistics, Universidad Nacional de Colombia
[2]apchavezr@unal.edu.co, Department of Statistics, Universidad Nacional de Colombia
[3]neacruzgu@unal.edu.co, Faculty of Nursing, Universidad Nacional de Colombia

A new methodology is presented to fit a surface model in a block survival problem, the estimation of the parameters involved in the proposed model is performed where the variable of interest is the time until the occurrence of an event. Theoretical development is carried out for the construction, estimation and validation of assumptions, using as a basis the Cox proportional hazards model and the response surface methodology. An adaptation of the classical tie correction methods is also developed for the proposed methodology. To evaluate the performance of the proposed model in comparison with other models, a simulation study is conducted to investigate the performance, properties and suitability of the model in previously specified scenarios. The results show that by combining the proportional hazard model with the response surface methodology, it is possible to identify the levels of treatments that optimize the response variable. A mortality study is carried out on grafted Inchi plants, considering the following factors: nitrogen doses (9, 12 and 18 gr of urea (46% N) plant-1/application) and time of removal of the plastic covering the graft (60, 90 and 120 days after grafting). In addition, the sex of the donor plant is taken into account as a natural block that eliminates the effects of unknown sources. Finally, this methodology has the advantage of being able to include the block which allows reducing the experimental error, improving efficiency by detecting differences between treatments, allowing more reliable comparisons between treatments.

Keywords: Proportional hazard, response surface methodology, block design, tie correction, time to event.

Beta regression model zero-inflated to measure the incidence of disease in tomato plants

Sandra Esperanza Melo Martínez[1], Oscar Orlando Melo Martínez[2] , Carlos Eduardo Melo Martínez[3]

[1] *semelom@unal.edu.co, Department of Agronomy/ Universidad Nacional de Colombia*
[2] *oomelom@unal.edu.co, Department of Statistics / Universidad Nacional de Colombia*
[3] *cmelo@udistrital.edu.co, Faculty of Engineering / Universidad Distrital Francisco José de Caldas*

In experiments in fields such as Agronomy and Biology, especially in the area of phytopathology, the incidence of disease in some crops is usually studied with repeated measurements presenting correlations between observations. Furthermore, this incidence is in the interval [0, 1) and often with too many zeros. In this work, a Beta regression model with random effects is fitted for an experimental design in blocks that was carried out in the municipality of Fomeque (Cundinamarca department, Colombia). This model is carried out to see the association between the incidence of botrytis in tomato plants using the independent variables: two fertilization plans of 6 and 8 kg per plant or treatments, in two locations Susa and Tablón, two axes of the plant and several flowering floors.

The model used includes two parts: a logistic regression component to model the presence and absence of the disease in plants, and a Beta regression component to model non-zero incidences. Each component includes a random effect to account for correlations among measurements made on the same plant. From the statistical analysis made with the model, no differences appeared between the fertility treatments, there are differences between the main and secondary axes for the variable incidence of botrytis cinerea in tomato flowers. Furthermore, it was found that axis two has a lower risk of presenting the disease and significant differences found between flowering floors. There are an increase in incidence over time between 84 and 140 days after transplantation with maximum between the floors 6 to 8.

Keywords: longitudinal Beta regression, zero inflated, incidence, random effect.

Age at menarche and its relationship to body mass index among adolescent girls in Chile: a joint modeling approach

Cristian Meza[1], *Danilo Alvares*[2], *Susana Eyheramendy*[3]

[1]cristian.meza@uv.cl, INGEMAT-CIMFAV, Universidad de Valparaíso, Chile
[2]danilo.alvares@mrc-bsu.cam.ac.uk, MRC Biostatistics Unit, University of Cambridge, UK
[3]susana.eyheramendy@uai.cl, Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Chile

Puberty is the period between childhood and adulthood when sexual development takes place. Menarche marks the beginning of the female reproductive capacity and is a significant developmental milestone for girls during this period. It typically occurs 2-2.5 years after puberty starts. Studies show that the average age at menarche has decreased from 17 years in 1840 to about 12 years in 2000 in most developed nations. The onset of menarche is influenced by both hereditary and environmental factors, with higher BMI being associated with earlier onset.

In this study, we analyze data from the Growth and Obesity Chilean Cohort Study to model the age at menarche based on BMI and other factors, modeling these variables jointly using a shared random-effects approach. We implemented a mixed model for the longitudinal marker, BMI, and a polynomial regression to model the age at menarche, including characteristics of the mixed model as covariates. Further, as an alternative approach, we categorize the age at menarche into three groups: early age menarche (<12 years), normal age menarche (between 12 and 14 years), and late age menarche (>14 years), and use a categorical data model.

Estimation of this kind of model may be challenging. Firstly, we propose to use a two-stage approach to obtain the maximum likelihood estimates based on linearization or approximation methods. As an alternative, we also apply the Stochastic Approximation version of the EM (SAEM) to estimate jointly the parameters of these models. With both approaches, we observe that the shared parameters seem to be significant which confirms the link between the BMI and the age at menarche.

Keywords: Joint model; Mixed effects models; SAEM algorithm

A bivariate zero-inflated Poisson regression model for partial body radiation exposures

Dorota Młynarczyk[1], *Pedro Puig*[2], *Carmen Armero*[3], *Virgilio Gómez-Rubio*[4]

[1]dorotaanna.mlynarczyk@uab.cat, Universitat Autònoma de Barcelona
[2]ppuig@mat.uab.cat, Centre de Recerca Matemàtica y Universitat Autònoma de Barcelona
[3]carmen.armero@uv.es, Universitat de València,
[4]virgilio.gomez@uclm.es, Universidad de Castilla-La Mancha

Keywords: biodosimetry, bivariate Poisson model, zero-inflated models

Bivariate zero-inflated Poisson models are a type of statistical models used to joint analyze two count variables that exhibit excess zeros and overdispersion. The bivariate Poisson model assumes that the two count variables are generated by two different Poisson processes, with a correlation term between them. In the bivariate zero-inflated model, the excess zeros in the count variables are accounted for by an additional process that generates zeros in both variables. These kind of models are commonly used in various fields, including epidemiology, ecology, and sports data [1]. This study explains the theoretical background of bivariate zero-inflated Poisson regression models and their application in biological dosimetry studies.

Biodosimetry is a method used to measure the radiation dose received by an individual from a radiation exposure. Dicentrics and chromosomal translocation are two of the most commonly used biomarkers to assess the amount of radiation exposure that an individual has received. These chromosomal aberrations occur when ionizing radiation damages the DNA in a cell, causing it to break and then rejoin with another piece of DNA, resulting in structural abnormalities in the chromosomes. The number of dicentrics and translocations in a person's cells can be used to estimate the amount of radiation exposure that they have received [2]. For example, higher levels of radiation exposure result in higher frequencies of dicentrics and translocations. Therefore the process of creating a calibration curve from the data provided before a possible radiation accident is important for the field of biodosimetry.

As dicentrics and translocations can be observed together when using fluorescence in situ hybridization (FISH) assay, we think it is beneficial to analyze both types of aberrations simultaneously. Partial body radiation exposures refer to situations where only certain parts of the body are exposed to ionizing radiation, while other parts are shielded from the radiation. This means that no irradiated cells have mostly zero observed aberrations, resulting in excess zeros in partial body exposure models, although some natural processes could also result in translocation formation. Taking all these thing into account we propose a novel strategy to deal with these kind of biodosimetric data using a bivariate zero-inflated Poisson regression model.

References

[1] Karlis, D., and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **52**, 381 – 393.

[2] Duran, A., et al. (2002). Suitability of FISH painting techniques for the detection of partial-body irradiations for biological dosimetry. *Radiation research*, **157**(4), 461 – 468.

Are green spaces contributing to gentrification in Valencia city?

Susana Morant-Garcia[1], *David V. Conesa*[2], *Francisco Palmí-Perales*[3]

[1]sumogar@alumni.uv.es, Dept. of Statistics and OR, Universitat de València
[2]david.v.conesa@uv.es, Dept. of Statistics and OR, Universitat de València
[3]francisco.palmi@uv.es, Dept. of Applied Economics, Universitat de València

The effect of the gentrification in urban areas has been the focus of several studies in the last decades. Although urban greening is universally recognized as an essential part of sustainable and climate-responsive cities, a growing literature on green gentrification argues that new green infrastructure, and greenspace in particular, can contribute to gentrification, thus creating social and racial inequalities in access to the benefits of greenspacie a nd further enviormental and climate injustice.

In this study we analyse the possible relationship of some social covariates and the green spaces on the gentrification among the city of Valencia. To do so we use a hierarchicial Bayesian approach to describe the gentrification index with respect the above mentioned covariates at censal area level. We consider autoregressive conditional approaches to analyse the spatial variability. This model is employed across different specific time periods. First results indicate that in Valencia there is no a clear pattern of green gentrification.

Keywords: Urban greening, Bayesian hierarchical models, Areal data

Longitudinal study on nutritional profile of oropharyngeal cancer patients according to HPV status: a challenge for the statisticians

F. Morey [1,2], *M. Choulli*[1,2,3,4], *R. Alvarez*[5,6], *X. Wang*[1,2], *B. Quirós*[2,8], *S. Tous* [2,8], *A.R. González-Tampán* [4,5], *M.A. Pavón*[1,9], *M. Gomà*[1,10], *M. Taberna*[1,5,7], *R. Mesia* [11], *L. Alemany* [1,2,8], *M. Oliva*[5,7], *M. Mena* [1,2,8]*, *L. Arribas*[1,4,5]*

[1]fmorey@idibell.cat, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain; [2]mchoulli@idibell.cat, Infections and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Barcelona, Spain; [3]University of Barcelona, Barcelona, Spain; [4]Clinical Nutrition Unit, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Barcelona, Spain; [5]Head and Neck Cancer Functional Unit, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Barcelona, Spain; [6]Medical Oncology Department, Canary University Hospital, San Cristobal de la Laguna, Tenerife, Spain; [7]Medical Oncology Department, Catalan Institute of Oncology (ICO), ONCOBELL, L'Hospitalet de Llobregat, Barcelona, Spain; [8]Centro de Investigación Biomédica en Red: Epidemiología y Salud Pública (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain; [9]Infections and Cancer Laboratory (INCALAB), Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Barcelona, Spain; [10]Pathology Department, Bellvitge University Hospital, L'Hospitalet de Llobregat, Barcelona; [11]Medical Oncology Department, Catalan Institute of Oncology (ICO)-Badalona, B-ARGO group, Barcelona, Spain. **(*)Both authors contributed equally as senior authors.**

Keywords: Human Papillomavirus (HPV), Oropharyngeal Cancer (OPC), longitudinal data.

In the last two decades, Human Papillomavirus (HPV) infection has emerged as a new causal agent for oropharyngeal cancer (OPC), resulting in a significant increase in the proportion of cases attributed to HPV in some world regions. HPV-related OPC patients (pts) have a better prognosis and longer survival compared to pts without the infection. On the other hand, malnutrition is a common problem in OPC pts. Initially nutrition impact symptoms, cause insufficient food intake compromising nutritional status and toxicity during oncological treatment, and can exacerbate malnutrition. The aim of this study was to assess and compare the nutritional status (NS) of pts regarding their HPV status, using nutritional parameters. At a statistical level, our aim was to understand and take into account the limitations of a longitudinal retrospective study.

A retrospective cohort of primary OPC pts treated with curative intent at the Catalan Institute of Oncology from 2016 to 2020 was recruited. We examined their nutritional, body composition and anthropometric parameters at three time points: at baseline, at three months and at six months after treatment (trt). Firstly, a sensitivity analysis was carried out to identify the bias caused by excluding pts because they did not have anthropometric data at baseline. Secondly, another sensitivity analysis was done to determine how the characteristic of the sample was affected by the losses in the follow-up. Finally, descriptive analysis was performed to explore the differences regarding HPV status of OPC pts at each time point and across the follow-up time. Sankey diagrams among nutritional support and assessment by HPV status were created to illustrate the dynamic flow of pts through time.

Initial recruited number of pts was 131. Pts with anthropometric data were 102. Sensitivity analysis comparing pts with and without anthropometric data showed a higher proportion of pts with initial stages (60.7% vs 34.3%, p=0.026) and surgery +/- CT/RT trt (51.7% vs 7.8%, p<0.001) in the excluded group. The number of pts with information at the baseline and at three months after trt was reduced at 71, observing that a higher proportion of pts with surgery +/- CT/RT trt continued being excluded.

Finally, the number of OPC pts included in the analysis was 71; 33 (46.5%) were HPV-related (HPV DNA and $p16^{INK4a}$ positive). HPV negative pts were significantly older, more smokers, drinkers than HPV-related OPC pts. Nutritionally, at baseline, HPV-related pts had a higher body mass index (BMI) (median: 27.3 vs 21.8 kg/m^2, p<0.001), and better NS (60.6% vs 26.3%, p=0.005). Conversely, after three months after trt, HPV-related pts reduced their BMI compared to baseline (median BMI for HPV-related: 27.3 at baseline vs 24.7 at three months, p=0.026, median BMI for HPV negative: 21.8 at baseline vs 20.6 at three months, p=0.223). At six months after trt, HPV negative and HPV-related OPC pts maintained similar BMI.

In conclusion, longitudinal data collected retrospectively involve a diversity of limitations that need to be clarified to correctly infer the results obtained. Longitudinal studies denote the importance of understanding the whole process: when, who, how and what information was collected for each patient to know if the selection is (and how) affecting the results. Only then, this knowledge can be helpful in the nutritional management of these pts. Nevertheless, more research is needed to better understand the changes observed.

Analysis of the Health-Related Quality of Life through PROreg R package: a case study of patients with eating disorders

Josu Najera-Zuloaga[1], *Dae-Jin Lee*[2], *Inmaculada Arostegui*[3]

[1]josu.najera@ehu.eus, Department of Mathematics University of the Basque Country UPV/EHU
[2] daejin.lee@ie.edu, IE University, School of Science and Technology, Madrid, Spain
[3] inmaculda.arostegui@ehu.eus, Department of Mathematics University of the Basque Country UPV/EHU; BCAM - Basque Center for Applied Mathematics, Bilbao, Spain

Patient-reported outcomes (PROs) are becoming important indicators of the health status of patients in observational and experimental studies. PROs are measured through questionnaires leading to several dimensions that constitute the PRO. The dimensions are constructed as a sum of ordinal responses to several items and, hence, they are defined as bounded and discrete scores, which, due to patients' perception variability, usually show U, J, or inverse J shapes. Consequently, PRO dimensions tend to have excess variability beyond the binomial distribution, a property called overdispersion.

In this context, beta-binomial distribution has been proposed in the literature to fit PRO dimensions, and beta-binomial regression as a good alternative for modelling purposes. However, the fact that the beta-binomial does not belong to the exponential distribution family limits its applicability in both distributional and regression framework.

In this work, we present the PROreg R-package, an innovative R package which offers a wide variety of functions that implement models based on the beta-binomial distribution. These functions can be can very useful when analysing overdispersed binomial data in regression framework, especially PROs such as Health-Related Quality of Life (HRQoL). With the aim of showing the contribution of the package to clinical application, we have developed two different model approaches to analyse the HRQoL of patients diagnosed with eating disorders who were followed up for two years at Galdakao-Usansolo Hospital in Biscay. The objective was to check the influence of clinical and sociodemographic variables on the HRQoL of the patients. In the first approach, we selected a specific HRQoL dimension and we aimed to analyse the effect of some covariates on the evolution of the dimension over time. Therefore, we applyed a longitudinal beta-binomial regression model, where specific matrices that define the temporal correlation within patients were defined. For the second approach, it is worth mentioning that in real practise, as it was done in the first approach, each of the dimensions that constitute the PRO is analysed separately. This approach can have a loss of information as PRO dimensions given by the same questionnaire tend to be correlated. Therefore, in this case, we applyed a multidimensional beta-binomial model to analyse the influence of clinical and sociodemographic variables in the dimensions of the HRQoL all together. Clinical interpretations of the results, such as the matrices that must be constructed before applying the functions that implement the models are explicitly shown.

Keywords: PROreg, Patient Reported Outcomes, Beta-binomial regression

Some linear models to biodiversity data from organic carbon in Puebla, Mexico

Oroza AA[1], *Grajales LF*[2], *Linares G*[3]

[1]aleyda16188@hotmail.com
[2]lfgrajalesh@unal.edu.co, Universidad Nacional de Colombia
[3]gladys.linares@correo.buap.mx

Climate change is an important problem in the world. The increasing amount of gases of the greenhouse effect within the atmosphere is the main cause of this change. One proposal to reduce these gases is storing up organic carbon (OC) in the soil by means of forest managing and woodlands conservation. Puebla, in Mexico, has rich ecosystems; they are kept as a special zones called *zonas terrestres prioritarias* (RTP's). Due to OC is an important element to measure the impact due to climate change in a region, in this work, we employed some mixed and random models in order to explain the OC in samples (litter), in Puebla, Mexico, RPT-105. Five *sites* or profiles (P30F, P35F, P36F, Pinhonero and Yucca) were analyzed; covariates include chemical and physical properties such as pH in water, pH in potassium chloride, *temperature* and *humidity*. Some of the results indicate that pH was an important covariate to explain the OC in a regression model, instead *temperature* and *humidity* were very important covariates to explain the OC in both fixed and mixed models ($p - values < 0.05$). Also, *sites* was a significative random effect to explain the organic carbon ($p - value < 0.05$).

Keywords: linear models, climate change, mixed linear models.

A Comparison of Infectious Disease Forecasting Combinations Methods

Manuel Oviedo-de la Fuente[1], *Rubén Fernández-Casal*[2], *José Antonio Vilar*[2]

[1]manuel.oviedo@udc.es, Department of Mathematics, CITIC, group MODES, University of A Coruña
[2], Department of Mathematics, CITIC, Research group MODES, University of A Coruña

Forecast combinations have prospered remarkably in recent years, being very useful for predicting the evolution of infectious diseases such as the flu or Covid-19, among many others. Within the framework of the Covid-19 pandemic and at the initiative of the Spanish Mathematics Committee (CEMat), a web tool was developed to show the evolution of indicators of interest in the pandemic together with short-term daily predictions of them (ForeCoop project, https://covid19.citic.udc.es). The predictions were actually cooperative predictions (meta-predictions), since they were built by combining predictions submitted by a wide range of research groups from all over Spain. Following these ideas, this work will try to compare ForeCoop implementation with other procedures available in R packages such as ForecastComb, Opera, ForecastCombinations and ForecastHybrid. Different multivariate and functional regression models will be used as individual predictors, trying to improve the accuracy of the predictions by integrating information collected from different sources (surveillance services, IoT platforms, social networks,...) and with different temporal or spatial level aggregation (day/week, region/country, other). Finally, aspects such as the automation of the process, CPU time, evaluation metrics, the difficulties inherent to the process and some conclusions about the usefulness of the work developed will be discussed.

Keywords: Forecast combinations, infectious diseases.

Dynamic zoning of agricultural plots based on satellite information

Paccioretti, P.[1], *Scavuzzo, M.*[2], *Balzarini M.*[3]

[1]pablopaccioretti@agro.unc.edu.ar, Comisión Nacional de Actividades espaciales. Universidad Nacional de Córdoba. Instituto Mario Gulich. Universidad Nacional de Córdoba, Facultad de Ciencias Agropecuarias, Departamento de Desarrollo Rural, Cátedra de Estadística y Biometría. Consejo Nacional de Investigaciones Científicas y Técnicas. Córdoba, Argentina.

[2]scavuzzo@conae.gov.ar, Comisión Nacional de Actividades espaciales. Universidad Nacional de Córdoba. Instituto Mario Gulich.

[3]monica.balzarini@unc.edu.ar, Unidad de Fitopatología y Modelización Agrícola (UFyMA). INTA-CONICET. Estadística y Biometría. Facultas de Ciencias Agropecuarias. Universidad Nacional de Córdoba. Argentina.

Supported by remote sensors information, farmers can identify key factors that impact crop yield and plan non-uniform agricultural management to make more efficient use of the natural and economic resources. In recent years, the availability of satellite products has grown, and it is possible to find multispectral images as well as synthetic aperture radar (SAR) images at a spatial scale that allows their use in the context of precision agriculture (PA). Based on this information, it is possible to derive dynamic indicators of crop status; however, the use of SAR-related information is not widely studied in PA. The objective of this study is to compare the efficiency of zoning agricultural plots using optical images, SAR images, or a combination of both at different crop stages. Images from five corn fields were analysed. For each field, images in six dates related to the crop cycle (V6, V10, V14, R1, R4) were obtained. All fields were zoned at each date, delimiting two to five homogeneous zones, using information from bands and indices derived from optical and radar images, and combining both sources. The zonings were performed using the KM-sPC method, that applies spatial principal components (sPC) on the satellite data and fuzzy k-means to cluster field sites using the sPC as input. In addition, a consensus cluster among zoning at different crop stages was performed for each agricultural field using the majority method. Differences between field zones, for each type of satellite data, were assessed by the pseudo-F index. It was observed that optical images performed better than radar images for zoning corn fields. In a few fields, SAR images provided extra information to improve zoning compared to optical images.

Keywords: remote sensing, precision agriculture, spatial principal components, fuzzy k-means

Statistical approaches for the integration of omics data

Carlos J. Peña[1], *Juan Antonio Carbonell*[2], *Sheila Zúñiga Trejos*[3]

[1]cpena@incliva.es, Biostatistics Unit, Biomedical Research Institute - INCLIVA
[2]jacarbonell@incliva.es, Biostatistics Unit, Biomedical Research Institute - INCLIVA
[3]smzuniga@incliva.es, Bioinformatics and Biostatistics Unit, Biomedical Research Institute - INCLIVA

Introduction: Combining multiple omics data sets can help provide a better understanding of diseases or complex biological processes. This combination may cover the data generated from genome, proteome, transcriptome, metabolome or epigenome but it can be further extended to other biological data.
The integration of individual omics data improve prognostics and predictive accuracy of disease phenotypes. In recent years, several tools have been developed for data integration that allow to identify subtypes of a disease based on multi-omics profiles (disease subtyping) or to predict biomarkers for diagnostics. In this work, we focus on some of the tools that perform integration of multiple omics data and we describe their methods.

Material: The Cancer Genome Atlas (TCGA) is a repository which includes genomic, epigenomic, transcriptomic, proteomic and clinical data for 32 cancers. We selected the Breast Invasive Carcinoma (BRCA) cohort and focused our analyses on characterizing the breast cancer subtypes.
For this tumor type, we downloaded TCGA-curated data sets containing gene expression, miRNA expression and protein abundance quantification as well as the patients' clinical information including the overall survival data. The individual omics data sets consist of mRNA expression, miRNA expression and protein abundance measured on 367 patients.

Methods: Comparative analysis was performed for four unsupervised multi-omics data integration methods: 1) multi-omic factor analysis [1] (implemented in the R package *MOFA2*); 2) regularized and sparse Generalized Canonical Correlation Analysis [4] (package *RGCCA* in the R software); 3) joint latent variable models implemented in the R package *iClusterPlus* [2]; and 4) Joint and Individual Variation Explained (JIVE) decomposition [3].

Future Work: We plan to evaluate the effect of the latent factors resulting from the proposed methods on overall survival along with other clinical features.

Keywords: data integration, multi-omics, disease subtyping.

[1] Argelaguet, R. et al. (2018) Multi-omics factor analysis- a framework for unsupervised integration of multi-omics data sets . *Mol. Syst. Biol*. **14**, e8124.
[2] Mo Q. et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci U S A*. **110**(11):4245-50. doi: 10.1073/pnas.1208949110.
[3] O'Connell MJ, Lock EF (2016). R.JIVE for exploration of multi-source molecular data. *Bioinformatics*. **32**(18):2877-9. doi: 10.1093/bioinformatics/btw324.
[4] Tenenhaus A. et al. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*. **15**(3):569-83. doi: 10.1093/biostatistics/kxu001.

Multivariate indices for experimental data: a comparative evaluation

Fabrícia Queiroz Mendes[1], *Marcelo da Silva Maia*[2], *André Mundstock Xavier de Carvalho*[2]
[1]fabricia.mendes@ufv.br, Institute of Agricultural Sciences, Federal University of Viçosa, Brazil
[2]Postgraduate Program in Agronomy, Institute of Agricultural Sciences, Federal University of Viçosa, Brazil

Multivariate statistical procedures applied to experiments can allow a better view of a whole set of response variables and allow adequate control of the type I error rates accumulated by experiment. In general, selection indices such as the Mulamba-Mock index and the Desirability index show robustness, simplicity and good discriminative power. Indices based on principal component analysis (PCA), however, may allow evidence of patterns of differences that selection indices cannot. The objective of this work was to empirically validate simple options for multivariate analysis, evaluating the accumulated empirical rates of type I error per experiment, and to compare the power of the multivariate indices based on PCA and the Mulamba-Mock, Desirability and its variations. Additionally, the objective was to validate simple modifications in the indices in order to increase their discriminative power and to evaluate objective criteria for selecting variables for assigning weights. The study was conducted based on the simulation of data from 1600 experiments, which were separated into four scenarios with 400 experiments each: experiments with 4 highly correlated response variables, under total nullity of effects for treatment (i); experiments with 8 response variables with the lowest level of correlation between them, also under total nullity (ii); experiments with 8 response variables with a lower level of correlation between them, but under partial null conditions (iii) and; experiments with 4 uncorrelated response variables, under total nullity of effects for treatments (iv). Six simple multivariate indices were calculated for each experiment: Mulamba-Mock index (MM), Desirability index (Di), Desirability index converted to rank scale (RT-Di), Desirability index without zero (Di-zm) and two proposals for modifying the Di-zm index (Modified Di-zm 1 and Modified Di-zm 2), in addition to two indexes based on PCA. In the simple multivariate indices, the type I error rates were also evaluated considering three criteria for attributing weights to one or two variables. The best results were obtained with the modified Di-zm index 1, where the power of the index reached 46% while the MM reached power of only 31.5% in the evaluated conditions. Attributing a lower weight to a variable, not defined a priori, increased power but resulted in a significant increase in the type I error rates of the simple indices when under zero covariance.

Keywords: Desirability index, principal component analysis, maximum family-wise error.

ICD9 to ICD10 update: effects in SVM fitting & prediction

Elies Ramon[1a], *Víctor Moreno*[2a,b]

[1]egurrea@idibell.cat, [a]Unit of Biomarkers and Susceptibility (UBS), Oncology Data Analytics Program (ODAP), Catalan Institute of Oncology (ICO). ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL). Centro de Investigación Biomédica en Red de Epidemiologia y Salud Pública (CIBERESP).

[2]v.moreno@iconcologia.net, [b]Department of Clinical Sciences, Faculty of Medicine and Health Sciences and Universitat de Barcelona Institute of Complex Systems (UBICS), University of Barcelona

Introduction: The International Classification of Diseases (ICD) is a worldwide healthcare classification system that codifies diagnoses and procedures encountered in medical settings. The 10th revision greatly expanded the number of codes of the 9th version, including 5 times as many diagnosis codes and 19 times as many procedure codes. Structural changes include slightly different arrangements of chapters and subchapters and a higher level of granularity and detail (3-5 characters for diagnosis codes in ICD9 vs 3-7 in ICD10). ICD9→10 translation is done with the General Equivalence Mapping (GEM) but, in most cases, there is not a one-to-one match between codes. Here, we analyse if the translation noise impacts the fitting and performance of machine learning models. Support Vector Machines (SVM) are especially suited for this task, as their dual form allows us to assess perturbations on the dataset as changes in the relationship between samples via the kernel matrices.

Materials & methods: The training set consisted of 9103 colorectal cancer cases and 9103 matched controls from the Catalan Institute of Healthcare (PADRIS database). For all patients, personal story of previous hospital admissions was coded in ICD9. We translated this data to ICD10 using GEM; if one-to-one match was not possible, one of the candidate ICD10 codes was chosen at random. Then we collapsed both datasets at the chapter, subchapter and category (3-code digit) level. For each level of granularity, we trained a linear SVM model using the original (SVM-9) and translated (SVM-10) data; we denote their respective kernel matrices as K and K'. The optimal value for the C hyperparameter was chosen with 5-Cross-Validation. Once we had the definitive model, the noise of the ICD9→10 translation was compared in five different ways: 1) Cosine similarity and relative spectral distance between K and K', 2) Stability of the support vectors in SVM-9 and 10, 3) Comparison of the feature importances recovered from the SVM-10 model and the SVM-9 model, 4) Accuracy of SVM-9 model fitting, and "refit" accuracy when switching the original K matrix for K' and 5) Prediction performance of both models in an independent test set (1602 cases and 7718 controls) coming from the same PADRIS database.

Results: Our original ICD9 data was grouped in 17 chapters, 129 subchapters and 845 categories, expanded to 20 chapters, 232 subchapters and 1221 categories in the ICD10 translation. Similarity between K and K' ranged between 75-92% and decreased with higher granularity. We observed the same pattern with support vectors (90-97% of SVM-9 also present in SVM-10). Prediction performance was modest (AUC≈0.59, weighted accuracy≈0.56) without significant differences between both models. Top most important features were also similar and included a history of previous neoplasms, anaemia and neurodegenerative or mental disorders. In summary, although translation slightly distorted the original kernel matrix, it did not have a strong effect in the fitting and prediction of the SVM.

Keywords: ICD, SVM, kernel matrix

Potential risk factors of injuries in professional football using Multivariate Survival Trees: a comparison of female vs. male football players

Jone Renteria[1], *Lore Zumeta-Olaskoaga*[1,2], *Eder Bikandi*[3], *Jon Larruskain*[3], *Dae-Jin Lee*[4]

[1]{jrenteria,lzumeta}@bcamath.org, Applied Statistics Research Line, Basque Center for Applied Mathematics, Bilbao, Bizkaia, Spain

[2]Departamento de Matemáticas, Universidad del País Vasco UPV/EHU, Leioa, Bizkaia, Spain

[3]{e.bikandi,j.larruskain}@athletic-club.eus, Athletic Club, Medical Services, Lezama, Bizkaia, Spain.

[4] daejin.lee@ie.edu, School of Science and Technology, IE University, Madrid, Spain

Nowadays, the use of information technologies has become a common and necessary practice among professional football clubs. During each match and training session, football teams collect a big amount of data from every movement that a particular player performs during the time he/she is on the field. Moreover, the medical team performs periodical check-ins that include performance physical tests, in which the players are screened for biomechanical movement, endurance, power, speed, and agility, among other characteristics. These screening tests are used to monitor players' functional and strength parameters, assess their physical fitness and thus provide crucial insight into areas that need improvement in order to reduce the likelihood of sports injuries.

In this work, we have gathered, cleaned, and curated all these multidimensional data to identify potential risk factors for sports injuries and better understand the event of injury by means of multivariate survival analysis for high-dimensional data. We have considered a Multivariate Survival Trees approach which can handle many covariates, requires few statistical assumptions, provides a sequence of prognostic rules that are easy to interpret, and performs nicely with the recurrent nature of sports injuries.

For the completion of this work, we have used combined multi-year data from five different seasons (from 17-18 to 21-22), for two female and three male senior teams from a Spanish professional football club. We have compared both sexes generating a different tree for each season. Finally, we have estimated survival curves for each terminal node of the trees, and we have ranked the relative importance of the covariates. Our method has been proven to be valuable for the identification of the potential risk factors involved in female and male non-contact lower limb injuries in football.

Keywords: sports injury, risk factor, multivariate survival tree

Dealing with Batch Effects in Metabolomics Data: A Comparison of ComBat, WaveICA2, and a Novel Residuals Method for Classification

Blanca Rius-Sansalvador[1], *Elies Ramon*[2], *Mireia Obón-Santacana*[3], *Victor Moreno*[4]

[1]brius@idibell.cat. Unit of Biomarkers and Susceptibility (UBS), Oncology Data Analytics Program (ODAP), Catalan Institute of Oncology (ICO)[a]. ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL)[b]

[2]egurrea@idibell.onmicrosoft.com [a,b]

[3]mireiaobon@iconcologia.net [a,b] Centro de Investigación Biomédica en Red de Epidemiologia y Salud Pública (CIBERESP)[c]

[4]v.moreno@iconcologia.net. [a,b,c] Department of Clinical Sciences, Faculty of Medicine and Health Sciences and Universitat de Barcelona Institute of Complex Systems (UBICS), University of Barcelona[d]

Background: Metabolomics analyses small molecules in biological samples, identifying biomarkers for diseases and effects of lifestyle factors. The identification of metabolomic differences between patients of different phenotypes is key to defining biomarkers for diagnosis. Unfortunately, large metabolomics studies are challenged by batch effects resulting from diverse sources. Non-biological systematic biases need identification and correction before biomarker discovery.

Purpose: Several methodologies have been proposed to correct batch effects of metabolomics data. We aim to identify the best one, that removes unwanted systematic differences between batches but preserves biological variability.

Methods: A study on plasma metabolomics data from 510 subjects (156 with colorectal cancer or high-risk adenoma, and 354 controls). The original data was transformed with arc-hyperbolic-sine to reduce asymmetry allowing for zero values. We propose a method based on the residuals of a linear model whose covariables are the Principal Components (PCs) where is it possible to identify clusters related to the batch. The residuals of this model are scaled with the original mean and standard deviation and used as new data. This method was compared to other published ones: ComBat and WaveICA2. Visual inspection of PC plots and temporal sequence plots were used to check the effect of the batch correction. The performance of a Random Forest classification model was used to assess variability relevant to the prediction of the cancer status.

Results: The representation of original data in a PC plot showed a striking batch effect. ComBat was a slightly worse method for removing batch effect in these data, since some grouping persisted. With WaveICA2 and the residuals method, samples did not cluster in the PC plots. The performance of Random Forest classification models, after an exhaustive hyperparameter optimization, showed similar cross-validated predictive AUC (around 0.70) for the method based on residuals and WaveICA2.

Conclusions: A simple method based on the residuals of a linear model adjusted for principal components was comparable to state-of-the-art WaveICA2 to reduce batch correction and maintain classification AUC.

Keywords: metabolomics, batch effect, residuals

Effectiveness and safety of tetanus vaccine administration by intramuscular vs. subcutaneous route in anticoagulated patients: Randomized cinical trial in primary care

Rodríguez Pastoriza, Sara[(1*)]*; Fernández Pérez*[(1)]*, Martín; Clavería Fontán, Ana*[(2,3)] *and Roca Pardiñas, Javier*[(3,4,5)]

[(1)] I-Saúde Group, Galicia Sur Health Research Institute, Vigo, Spain.
[(2)] Epidemilogist, Quality and Research Unit, EOXI Vigo, Galician Health Service, RedIAPP.
[(3)] Research Network on Chronicity, Primary Care and Health Promotion (RICAPPS), Vigo, Spain.
[(4)] Department of Statistics and Operations Research, University of Vigo, Vigo, Spain.
[(5)] Center for Research and Mathematical Technology of Galicia (CITMAga), Santiago de Compostela, Spain.

[*] sara.rodriguez@iisgaliciasur.es

BACKGROUND

Although tetanus is a disease of low incidence in Spain, it is an important public health problem due to the high associated mortality rate. Annual reports of registered cases show a gradual fall, with an average of 10 cases per year between 2009 and 2015, resulting in 18.4% being lethal.

METHODS AND MATERIALS

Design and study population

Prospective, double-blind clinical trial comparing tetanus-diphtheria vaccine administration routes, intramuscular (IM) vs. subcutaneous (SC) injection, in patients with oral anticoagulants. ISRCTN69942081. Patients treated with oral anticoagulants, 15 health centers, Vigo (Spain). Sample size, 117 in each group.

Outcome variables

Safety analysis: systemic reactions and, at the vaccine administration site, erythematic, swelling, hematoma, granuloma, pain. Efectiveness analysis: differences in tetanus toxoid antibody titers. Independent variables: route, sex, age, baseline serology, number of doses administered.

Analysis

We conducted a descriptive study of the variables included in both groups (117 in each group) and a bivariate analysis. Fewer than 5% of missing values. Imputation in baseline and final serology with the median was performed. Lost values were assumed to be values missing at random. We conducted a descriptive study of the variables and compared routes. For safety,multivariate logistic regression was applied, with each safety criterion as outcome and the independent variables.Odds ratios (ORs) were calculated. For effectiveness, a generalized additivemixedmodel, with the difference between final and initial antibody titers as outcome.

RESULTS

A previously published protocol was used across the 6-year study period. The breakdown by sex and route showed: 102 women and 132 men; and 117 IM and 117 SC, with one dose administered in over 80% of participants. There were no differences between groups in any independent variable. The second and third doses administered were not analyzed, due to the low number of cases. In terms of safety, there were no severe general reactions. Locally, significant adjusted differences were observed: in pain, by sex (male, OR: 0.39) and route (SC, OR: 0.55); in erythema, by sex (male, OR: 0.34) and route (SC, OR: 5.21); and in swelling, by sex (male, OR: 0.37) and route (SC, OR: 2.75). In terms of effectiveness, the model selected was the one adjusted for baseline serology.

A simplified model for the characterization of blood alcohol elimination

M.T. Santos Martín[1], *J.M. Rodríguez Díaz*[2], *I. Mariñas del Collado*[3]

[1] maysam@usal.es , Department of Statistics, University of Salamanca

[2] juanmrod@usal.es, Department of Statistics, University of Salamanca

[3] marinasirene@uniovi.es, Department of Statistics and Operational Research and Didactics of Mathematics, University of Oviedo

Abstract:

The model typically used to describe the elimination of alcohol concentration in humans assume a zero-order kinetics, that is, the alcohol is eliminated with a constant rate. However, it does not consider the absorption phase in which the alcohol concentration increases to reach a certain maximum peak. Some alternative models including both phases have been already studied, among them are the compartmental systems and the gamma distribution model.

In this work, a simplified model like the last one is proposed in order to be easy-to-use for the non-statistician community, such as forensic scientists. Optimal designs have been computed for this model, and compared with existing designs found in literature. In addition, equally spaced designs were studied, since this type of designs are usually preferred by experimental practitioners who would rather to take samples covering the whole design interval instead of restricting themselves to the fewer points contained in the optimal designs. Efficiencies of these designs with respect to the optimal ones have also been computed. In this setup, it will be necessary to assume a covariance structure between responses, however the case of independent observations has also been considered in order to be able to compare the results with those in works existing in literature.

Keywords: Optimal design of experiments, D-optimality, ethanol, Widmark

A new testing procedure for determining groups in cumulative incidence curves

Marta Sestelo[1], *Luis Meira-Machado*[2], *Nora M. Villanueva*[3] *and Javier Roca-Pardiñas*[4]

[1]sestelo@uvigo.es, CITMAga, Santiago de Compostela, Spain, Department of Statistics and O.R and SiDOR Group, University of Vigo

[2]lmachado@math.uminho.pt, Centre of Mathematics and Department of Mathematics, University of Minho - School of Sciences, Guimarães, Portugal

[3]nmvillanueva@uvigo.es, Department of Statistics and O.R., SiDOR Group and CINBIO, University of Vigo, Spain

[4]roca@uvigo.es, CITMAga, Santiago de Compostela, Spain, Department of Statistics and O.R and SiDOR Group, University of Vigo

The cumulative incidence function is the standard method for estimating the marginal probability of a given event in the presence of competing risks. One basic but important goal in the analysis of competing risk data is the comparison of these curves, for which limited literature exists. We proposed a new procedure that lets us not only test the equality of these curves but also group them if they are not equal. Note that, by clustering the cumulative incidence functions, we can identify subgroups of individuals who have different probabilities of experiencing each of the competing events and this can be useful for understanding the heterogeneity of the population and tailoring interventions to specific subgroups.

The proposed method allows determining the composition of the groups as well as an automatic selection of their number. Simulation studies show the good numerical behaviour of the proposed methods for finite sample size. The applicability of the proposed method is illustrated using real data.

Keywords: Clustering; Competing risks; Cumulative incidence function;

Enhancing Urban Public Transportation Efficiency through Accurate Passenger Volume Prediction: A Bayesian Spatial-Temporal Model Applied to Beijing Metro

He Sun[1], *Stefano Cabras*[2]

[1]sunhe1509@gmail.com, Department of Statistics, University Carlos III of Madrid and Beijing Metro Group Ltd (China)

[2]stefano.cabras@uc3m.es, Department of Statistics, University Carlos III of Madrid

With the purpose of diminishing the carbon emission, it is important to increase the efficiency of city public transportation and to do this, we have to predict the number of passengers. It is imperative to enhance the efficiency of urban public transportation systems. A critical component of this objective is the accurate prediction of passenger volume, which facilitates improved planning and resource allocation. This research introduces a Bayesian spatial-temporal model designed to forecast station occupancy in metropolitan subway transportation systems, thereby contributing to the reduction of traffic congestion and, consequently, the city's ecological footprint.

The proposed model not only yields precise point estimations of daily passenger flow but also provides a robust assessment of the associated uncertainty. This information enables a comprehensive understanding of traffic patterns, ultimately facilitating the development of more efficient public transportation networks. By optimizing these networks and decreasing reliance on private vehicles, this approach contributes to the reduction of carbon emissions and promotes ecological benefits.

In a practical context, the model has exhibited prediction accuracy that aligns with the standards set forth by the Beijing Metro enterprise. It is currently employed by Beijing Mertro Group Ltd to refine daily train schedules, exemplifying the potential of such models in fostering environmentally-conscious, efficient urban transportation systems that benefit both the environment and urban populations.

Keywords: Carbon emission, Bayesian spatial-temporal model, Public transportation efficiency.

Applied statistics as an essential tool for the success of the relationship between epidemiology and clinics: the study of the involvement of Human Papillomavirus with oropharyngeal cancer.

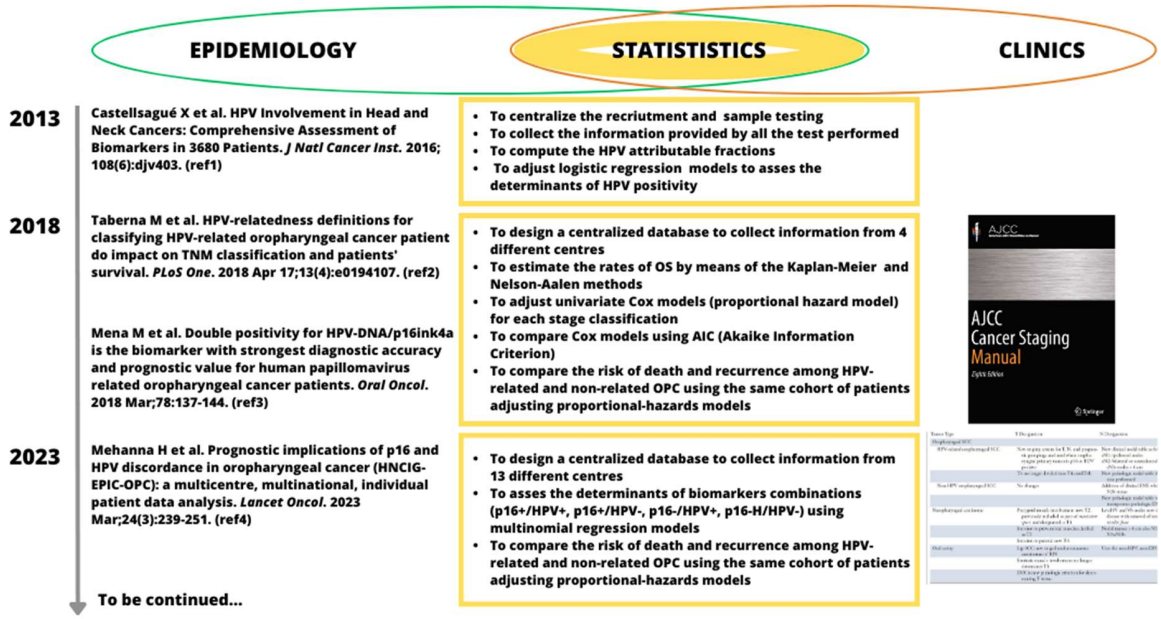
Sara Tous[1,2], Miren Taberna[3], Marisa Mena[1], Beatriz Quirós[1,2], Francisca Morey[1], Hisham Mehanna[4], Laia Alemany[1,2]

[1]stous@iconcologia.net. Programa de Recerca en Epidemiologia del Càncer, Unitat d'Infeccions i Càncer - Molecular, Institut Català d'Oncologia (ICO)-Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Spain; [2]Centro de Investigación Biomédica en Red: Epidemiología y Salud Pública (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain; [3]Departament d'Oncologia Mèdica, ICO-IDIBELL, L'Hospitalet de Llobregat, Spain; [4]Institute of Head and Neck Studies and Education (InHANSE), Institute of Cancer and Genomics Sciences, University of Birmingham, Birmingham, UK.

Keywords: Human Papillomavirus (HPV), Oropharyngeal Cancer (OPC), Cohort data analysis.

A multidisciplinary team composed by epidemiologists, statisticians, pathologists, and laboratory technicians at ICO started around 10 years ago to study the relationship between Human Papillomavirus (HPV) infection and oropharyngeal cancer (OPC). A large international study[ref1] including 3680 samples was conducted to estimate fractions (AF) of head and neck cancers (HNCs) attributable to HPV using six biomarkers. We observed that HPV contribution to HNCs was substantial but highly heterogeneous by cancer site, region, and sex, and confirmed the important role of HPVs in OPC. In 2018, we started to collaborate with the oncologist's team from the hospital given the different nature and better outcomes of OPC associated with HPV infection they were observing in the clinical practice. The etiologic role of HPV in OPC was well established at that time point. Nevertheless, information on survival differences by anatomic sub-site or treatment remained scarce. Simultaneously, a novel clinical stage classification for HPV-related OPC was just accepted for HNCs tumors classification, based on p16INK4a (p16) detection. However, it was still unclear the HPV-relatedness definition with best diagnostic accuracy and prognostic value. So, we conducted several studies to determine in a cohort of patients (pts) from Barcelona which could be the best definition to classifying HPV-related OPC pts[ref2] and to assess the determinants of HPV infection and prognostic values of OPC pts based on p16 and HPV detection[ref3]. We observed that HPV-relatedness definition does impact on TNM classification and the survival of p16+/HPV- pts was worse than p16+/HPV+. So, we extended our research to a multicenter study collecting multinational individual pt data including retrospective cohorts of consecutively recruited OPC pts previously analyzed[ref4]. The study included 7654 OPC pts from 13 different centers. We identified significantly different proportion of p16+/HPV- pts by geographical region, being highest in the areas with lowest HPV-AFs (r=–0.7, p=0.003). 5-year overall survival was different depending on p16/HPV detection: 81.1% (95% CI 79.5–82.7) for p16+/HPV+, 40.4% (38.6–42.4) for p16-/HPV-, 53.2% (46.6–60.8) for p16-/HPV+, and 54.7% (49.2–60.9) for p16+/HPV-, and the prognosis of discordant p16+/HPV- tumors also differed on smoking status. In conclusion, pts with discordant OPC (p16-/HPV+ or p16+/HPV-) had a significantly worse prognosis than pts with p16+/HPV+ OPC, and a significantly better prognosis than pts with p16-/HPV- OPC. Along with routine p16, HPV testing should be mandated for clinical trials for all pts. In Figure 1 we detail the contribution of the statistician in each study.

Figure 1: Timeline of the studies conducted and the role of the statistician.



CEB
EIB 2023

XIX Conferencia Española y VIII Encuentro Iberoamericano de Biometría

Vigo, del 27 al 30 de junio de 2023

Universidade Vigo

Departamento de Estatística e Investigación Operativa

SiDOR
Statistical Inference, Decision & Operations Research Group

Escola de Enxeñaría Industrial

CONCELLO DE VIGO

DEPUTACIÓN PONTEVEDRA

SEB Sociedad Española de Biostadística

IBS