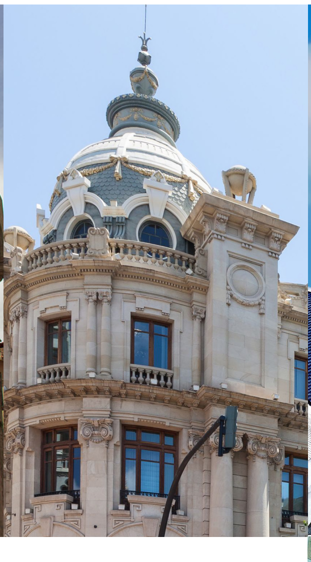


CEB 2019 EIB

Libro de Resúmenes

XVII CONFERENCIA ESPAÑOLA DE BIOMETRÍA
VII ENCUENTRO IBEROAMERICANO DE BIOMETRÍA

València, 19-21 de Junio de 2019

Diseño: 27P8 Studio.
Fotografías: Osho 8

Libro de resúmenes

XVII Conferencia Española y
VII Encuentro Iberoamericano de Biometría
CEB-EIB 2019

Valencia, 19-21 de junio

Libro de resúmenes **XVII Conferencia Española**
y VII Encuentro Iberoamericano de Biometría CEB-EIB 2019
ISBN: 978-84-09-11706-2

Equipo Editorial:

C. Armero Cervera, *Universitat de València*

D.V. Conesa Guillén, *Universitat de València*

A. Forte Deltell, *Universitat de València*

M. Trottini, *Universidad de Alicante*

A. M. Debón Aucejo, *Universitat Politècnica de València*

Comité Científico

Carmen Armero [Presidenta], *Universitat de València*

Inmaculada Arostegui, *Universidad del País Vasco UPV/EHU*

David Conesa, *Universitat de València*

María Durban, *Universidad Carlos III de Madrid*

Susana Eyheramendy, *Pontificia Universidad Católica de Chile*

Lupe Gómez, *Universitat Politècnica de Catalunya*

Dolores Jiménez, *Universidad de Sevilla*

Klaus Langohr, *Universitat Politècnica de Catalunya*

Raul Macchiavelli, *Universidad de Puerto Rico*

Miguel Ángel Martínez-Beneito, *Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana*

Vicente Núñez-Antón, *Universidad del País Vasco UPV/EHU*

Pere Puig, *Universitat Autónoma de Barcelona*

María del Carmen Romero, *Universidad Nacional del Centro de la Provincia de Buenos Aires*

Omar Ruíz-Barzola, *Escuela Superior Politécnica del Litoral (ESPOL)*

Lola Ugarte, *Universidad Pública de Navarra*

COMITÉ ORGANIZADOR

Anabel Forte [Presidenta], *Universitat de València*

Xavier Barber, *Universitat Miguel Hernández de Elche*

Paloma Botella-Rocamora, *Conselleria de Sanitat Universal i Salut Pública*

David V. Conesa, *Universitat de València*

Ana Corberán, *Universitat de València*

Ana María Debón Aucejo, *Universitat Politècnica de València*

Olga Susana Filippini, *Universidad de Buenos Aires*

Luis Fernando Grajales Hernandez, *Universidad Nacional de Colombia*

María Victoria Ibáñez Gual, *Universitat Jaume I*

Yasna Orellana, *Universidad de Chile*

Francisco Santonja, *Universitat de València*

Mario Trottini, *Universidad de Alicante*

Jessica Martha Vera, *Escuela Superior Politécnica del Litoral, Ecuador*

PATROCINADORES:

Universitat de València

Universitat Politècnica de València

Universitat d'Alacant
Universidad de Alicante

Universitat Jaume I

Institut Universitari de Matemàtiques i Aplicacions de Castelló · IMAC

Universitas Miguel Hernández

Generalitat Valenciana
Conselleria d'Educació, Investigació, Cultura i Esport

VAlencia BAyesian Research group

CAMPUSHABITAT5U
Campus de Excelencia Internacional

SEB
IBS
Región Española de la IBS
Spanish Region of the International Biometric Society

CIO
Universitas Miguel Hernández
Research Institute

SYNTAX®
FOR SCIENCE

SORT
Statistics and Operations Research Transactions

Corría el año 1995 cuando la *Conferencia Española de Biometría* (CEB) aterrizaba por primera vez en València. Era su 5ª edición y presidía el comité organizador Emilio Carbonell quien, dos años más tarde, pasaría a ser presidente de la Sociedad Española de Biometría (SEB). En esta ocasión la CEB vuelve a València para celebrar su decimoséptima edición y lo hace de la mano del departamento de Estadística e Investigación Operativa de la Universitat de València, con la colaboración de todas las universidades públicas valencianas y de la Generalitat Valenciana. Además, como cada cuatro años, la CEB, se une al *Encuentro Iberoamericano de Biometría* (EIB) en el que, por primera vez en ésta su séptima edición, se unen los esfuerzos de cinco regiones de la International Biometric Society (IBS): Argentina, Centroamérica y Caribe, Chile, Ecuador y España.

Edición tras edición, la CEB-EIB se ha convertido en un encuentro entre colegas de profesión que han acabado convirtiéndose en amigos y amigas y conformando una comunidad, nacida del trabajo, el esfuerzo y la colaboración. Una comunidad que se nutre y crece gracias a jóvenes investigadores e investigadoras cuya importancia no escapa a nadie. Es por ello que tanto la SEB como la IBS han apostado, un año más, por fomentar su participación lo que se refleja tanto en la cantidad como en la calidad de inscripciones de estudiantes y recién doctorados/as. Repetimos, además, la sesión "Young Statistician Showcase" que contará con 5 ponencias de jóvenes, seleccionadas únicamente en base a criterios científicos, entre todas las que presentaron un resumen extendido de su trabajo.

En este sentido, quisiéramos agradecer el magnífico trabajo del Comité Científico por su ardua labor a la hora de seleccionar las ponencias de las que podremos disfrutar. En especial, agradecer a Carmen Armero su labor de consenso y su disponibilidad para encontrar sinergias y que ambos comités pudiesen avanzar en sintonía. Agradecemos también la participación de los profesores Klaus Langohr y Adrian Bowman, y de la profesora Michela Cameletti que nos ofrecerán conferencias plenarias que seguro serán muy interesantes para todos. Y no podemos olvidar nuestro agradecimiento al profesor Virgilio Gómez Rubio por impartir el curso que da el pistoletazo de salida a esta conferencia.

Finalmente, debemos agradecer a todas las personas que han colaborado con este comité, en especial a nuestros/as Ayudantes y a las técnicas de ADEIT, así como a los patrocinadores que han hecho posible que podamos ofreceros el mejor ambiente científico en el mejor entorno posible. Esperemos que lo disfrutéis tanto como nosotros y nosotras lo hemos hecho preparándolo.

¡Bienvenidos a la CEB-EIB 19! ¡Bienvenidos a València!

Comité Organizador CEB-EIB 2019

Índice General

Conferencias Plenarias

Statistics with a human face

A. Bowman[1]

adrian.bowman@glasgow.ac.uk

[1] School of Mathematics and Statistic, The University of Glasgow

Human faces are of intrinsic interest to us all but the study of facial shape also has many biological and anatomical applications. For example, these include assessing the outcome of facial surgery and investigating the possible developmental origins of some adult conditions. Advances in accessible forms of 3D imaging are making this kind of data much more accessible. For analysis, an initial challenge is to structure the raw images by identifying features of the face. Ridge and valley curves provide a very good intermediate level at which to approach this, as these provide a good compromise between informative representations of shape and simplicity of structure. Some of the issues involved in analysing data of this type will be discussed and illustrated. Modelling issues include simple comparison of groups, the measurement of asymmetry and longitudinal patterns of shape change. This last topic is relevant at short scale in facial animation, medium scale in individual growth patterns, and very long scale in phylogenetic studies.

Statistical challenges in air pollution health risk assessment

M. Cameletti[1]

michela.cameletti@unibg.it

[1] Department of Management, Economics and Quantitative Methods, University of Bergamo

Air pollution remains one of the major environmental problems in Europe, affecting health and well-being of European citizens. Assessment of the impacts of air pollution on population health requires regularly updated and accurate exposure estimates at a proper spatial resolution.

Studying the association between health and air pollution through ecological regression models poses some statistical challenges. Firstly, data regarding pollutant concentration and health outcomes (i.e. mortality/morbidity data) are available from different sources, can be measured with error and are usually spatially misaligned. This means that pollutant concentration has to be upscaled at the area level (i.e. the change of support problem), while being measured at a finite number of point-referenced monitoring stations or simulated at the grid level by dispersion models. The second statistical challenge concerns how to link the exposure with the health outcomes taking into account all the uncertainty sources related to the air pollution field, in order to have unbiased estimates of the risk effect associated to air pollution exposure.

In my presentation I will discuss these challenging issues while implementing a model for air pollution health risk assessment, both through a simulation study and a real case application regarding hospitalization data in Italy. I will also present a space-time model to integrate spatially misaligned air pollution data coming from multiple data sources presenting the results regarding the case-study of NO2 concentration in Greater London.

Analysis of interval-censored data. Methods, software, and real world examples

K. Langohr[1]

klaus.langohr@upc.edu

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

In time-to-event studies, interval-censored data are encountered whenever the occurrence of the event of interest cannot be observed exactly but only within a time window. Medical examples for such so-called silent events are HIV infection or an immunological markers exceedance over a certain threshold.

The adaptation of methods for right-censored data to the analysis of interval-censored data is not straightforward, because for times within the observed intervals, the number of individuals at risk for the event of interest cannot be determined exactly. For this reason, in the last decades, many specific methods have been developed to deal with interval-censored data and, in parallel, a large number of R packages that implement these methods have been published.

In this talk, we will give an overview of existing methods for the analysis of intervalcensored data including nonparametric methods, parametric and semiparametric survival models, multi-state models, and regression models with interval-censored covariates. The methods and their implementation in R will be illustrated with real world examples from studies on the shelf life of foods, the evolution of bone mineral density in HIV-infected patients, and HIV vaccination studies.

Keywords: HIV studies, interval censoring, R packages

Conferencias Invitadas:
Sesión Iberoamericana. Jhonny Demey

A Control chart based on a Negative Binomial model to detect irregular values in infant mortality in Ecuador

Sandra García-Bustos[1], *Ana Debón*[2], *Iris Bustamante*[3]

[1]slgarcia@espol.edu.ec, Facultad de Ciencias Naturales y Matemáticas, Escuela Superior Politécnica del Litoral.

[2]andeau@eio.upv.es, Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València

[3]ijbustam@espol.edu.ec, Facultad de Ciencias Naturales y Matemáticas, Escuela Superior Politécnica del Litoral.

In this study, control charts have been designed to detect anomalies in the data of infant mortality in Ecuador. Control charts are based on hypothesis tests for changes in the parameters of a population and they are usually constituted by a control statistic and control limits. To design them, we have used regression models based on Poisson, Binomial and Negative Binomial distributions, considering factors such as gender, area and their interaction. The overdispersion of the data was better explained by the Negative Binomial Regression model, with all the explanatory variables being statistically significant. Based on the aforementioned, Negative binomial regression models were fitted for the infant mortality rate considering the interaction between gender and the urban and rural areas of Ecuador, that is, four models were obtained in the period from 1990 to 2017. The control chart used is based on an EWMA (Average of exponential weighting) version for the Pearson residuals of each of the four models. The application of these charts to mortality in the countries of Latin America is suitable since they are very efficient to detect small shifts in the means of the variables that are monitored. First, Phase 1, Construction of the control chart was carried out. In this phase, to determine the statistics and control limits, the EWMA chart for the Pearson residuals with a weight of 0.05 is constructed and an approximate confidence level of 95% is obtained for the four models. In phase 1, some Pearson residuals were out of the control limits, then these residuals were removed, and the control limits were readjusted until all the points were within the control limits. Finally, in Phase 2 of Monitoring, these statistics and control limits are proposed to control the future mortality rate for children in urban and rural areas in Ecuador.

Keywords: Pearson residuals, Binomial Negative regression, EWMA.

Goodness of fit for models with intractable likelihood

María Eugenia Castellanos[1], *Stefano Cabras*[2], *Oliver Ratmann*[3]

[1]maria.castellanos@urjc.es, Universidad Rey Juan Carlos (Spain) and Universitá degli Studi di Cagliari (Italy)

[2]stefano.cabras@uc3m.es, Universidad Carlos III de Madrid (Spain) and Universitá degli Studi di Cagliari (Italy)

[3]oliver.ratmann@imperial.ac.uk, Imperial College London (UK)

In recent years many statistical applications involve stochastic models with analytically intractable likelihood functions in areas as genetics, epidemiology or population biology, just to mention some. The rapidly growing literature on Approximate Bayesian Computation (ABC) has led to a set of methods which do not involve direct calculation of the likelihood, leading to approximate Bayesian inference for unknown parameters. In this work, we analyze the problem of checking the compatibility of a proposed stochastic model with the observed data. In the context of non alternative models, Bayes factors are precluded and only measures of 'surprise', such as p-values could be used. Here, we show that, even for models whose likelihood is not available in a closed form expression, calibrated conditional predictive p-values can be efficiently obtained as a by-product of ABC without any additional computational cost. We show that these are calibrated, that is, asymptotically uniformly distributed in [0,1] under the null hypothesis that the data are generated from the posited model, assuming general conditions on the summary statistics. The technique is illustrated on analytically tractable examples and on a complex tuberculosis transmission model.

Keywords: Approximate Bayesian Computation, model adequacy, model checking, simulation-based modelling.

AMS: 62F15, 62F03

Bayesian antedependence model proposals for longitudinal data

Edilberto Cepeda-Cuervo[1], *Vicente Núñez-Antón*[2]

[1]ecepedac@unal.edu.co, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá

[2]vicente@nunezanton@ehu.eus, Departamento de Econometría y Estadística (E.A. III), Universidad del País Vasco UPV/EHU, Bilbao

An important problem in Statistics is the study of longitudinal data taking into account the effect of other explanatory variables such as treatments and time and, at the same time, incorporate into the model the time dependence between observations on the same individual. The latter is specially relevant in the case of having nonstationary correlation, as well as nonconstant variance for the different time point at which measurements are taken. Antedependence (AD) models constitute a well known commonly used set of models that can accommodate this behavior. These covariance models can include too many parameters and estimation can be a complicated optimization problem requiring the use of complex algorithms and programming. In this paper, a new Bayesian approach for analyzing longitudinal data within the context of antedependence models is proposed. This innovative approach takes into account the possibility of having nonstationary correlations and variances, and proposes a robust and computationally efficient estimation method for this type of data. We consider the joint modelling of the mean and covariance structures for the general AD model, estimating their parameters in a longitudinal data context. Our Bayesian approach is based on a generalization of the Gibbs sampling and Metropolis-Hastings by blocks algorithm, properly adapted to the AD models longitudinal data settings. Finally, we illustrate the proposed methodology by analyzing several examples where AD models have been shown to be useful: the small mice, the speech recognition and the race data sets.

Keywords: Antedependence models, Bayesian methods, nonstationary correlation.

AMS: 62F15, 62J05.

Statistical challenges of present plant breeding programs

Julio A. Di Rienzo[1]

[1]dirienzo@agro.unc.edu.ar, Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, Argentina.

The basic data of a plant breeding program consists of three pieces of information: yield, and two labels: one identifying the genotype, the other, the environment where the yield was obtained. This information is intended to answer the WWW question (who won where?). This basic information is available in every breeding program database, and was, during a long time, the primary source for decision making. The days of getting 2-5% annual increase of yield, due to selection, have gone. Present efforts are handling a 1% annual gain target, including for transgenic materials. Achieve superior yields is intrinsically difficult because yield is a trait with a low heritability, which means that most of its variance is explained by environment conditions. To better understand the yield variation and make better agronomic management recommendations as well as to update the regionalization of the breeding programs, nowadays plant breeding programs are expanding the labelling part of the basic information. New dimensions of labelling include information at different levels. At genotype level there is an increasing number of molecular markers available for parent as well as for their crosses. At environmental level the layers of information are constantly increasing and cover very different scales and type of information. From satellite data monitoring crop changes during the growing season, to data at plot level describing soil characteristics or plants physiological status. This information is generated and used at different steps and in different ways within a breeding program. The statistical and the data sciences challenge, is to integrate the huge number of resulting covariates, typically highly correlated, to explain the results of the multi-environments trials. Classical mixed linear model approach is reaching its limit under this scenario, mainly due to the increasing dimensionality of the covariates space and the huge correlation among them. Under this scenario, the linear model approach has a poor predictive value, no to say that frequently the models can't be even estimated. So, even known methods as Random Forest, PLS, PCA (just to name some of them), and applied in creative ways, incorporating geographic information into their algorithms are gaining space in the statistical practices applied to plant breeding. Nowadays statisticians, involved in plant breeding programs, must develop original ways to produce new decision-making information, from the diverse and increasingly amount of data available from many sources at very different temporal and special scale.

Keywords: Integrating metadata, plant breeding.

The segmentation of series with a functional effect: a Bayesian approach

Cristian Meza[1], *Meili Baragatti*[2], *Karine Bertin*[3], *Emilie Lebarbier*[4]

[1]cristian.meza@uv.cl, CIMFAV-Facultad de Ingeniería, Universidad de Valparaíso
[2]meili.baragatti@supagro.fr, Montpellier SupAgro, MISTEA
[3]karine.bertin@uv.cl, CIMFAV-Facultad de Ingeniería, Universidad de Valparaíso
[4]emilie.lebarbier@agroparistech.fr,INRA, AgroParisTech

In some application fields, series are affected by two different types of effects: abrupt changes (or change-points) and functional effects. We propose here a Bayesian approach that allows us to estimate these two parts. Here the underlying piecewise-constant part (associated to the abrupt changes) is expressed as the product of a lower triangular matrix by a sparse vector and the functional part as a linear combination of functions from a large dictionary where we want to select the relevant ones. This problem can thus lead to a global sparse estimation and a Stochastic Search Variable Selection approach is used to this end. Our estimation method is based on MCMC algorithms (Metropolis-Hastings algorithm and Gibbs sampler). Although these algorithms can take more time than those used in a frequentist approach, our procedure benefits from the Bayesian framework, which results in two important aspects. The first one is that posterior distributions of the parameters are obtained. From these distributions, different quantities can easily be derived as credibility intervals of the means, the change-points and the selected functions, or the probability to have a change-point in a given interval. The second important aspect is that we can introduce expert knowledge through prior distributions. The performance of our proposed method is assessed using simulation experiments. Applications to two real datasets from environmental research and agronomy are also presented. More specifically, we propose to use our procedure in the geodesic field for the problem of homogenization of GPS series and to analyse the Périgord black truffe production in France.

Keywords: Segmentation,functional effect, Bayesian inference.

Conferencias Invitadas: Sesión Sociedad Española de Epidemiología

Bayesian shared component models for incidence predictions based on mortality

Jaione Etxeberria[1,2,3,*], *Tomás Goicoa*[1,2,4],*María Dolores Ugarte*[1,2,5]

[1] Department of Statistics, Computer Science and Mathematics, Public University of Navarre

[2] Institute for Advanced Materials (InaMat), Public University of Navarre

[3] Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP)

[4] Research Network on Health Services in Chronic Diseases (REDISSEC)

[5] Centro Asociado de Pamplona, UNED

[*] email: jaione.etxeberria@unavarra.es

For a proper allocation of health resources in cancer, different indicators such as incidence and mortality rates and counts are taken into account. In Spain cancer mortality figures are routinely recorded by Statistical Offices while incidence figures are systematically recorded by regional cancer registries. Generally, due to administrative procedures, incidence numbers become available three or four years later than mortality figures. In this context, to predict incidence rates in periods when the mortality is already known becomes necessary in order to provide the most updated cancer description. According to International Cancer Agencies, realistic predictions of incidence rates should fulfill a list of requirements: 1-They should be stable over time, 2-They must be comparable in different populations or regions, 3-Age-specific incidence curves should be provided (including childhood cancer rates) and 4-Mortality-to-Incidence ratios should be taken into account. Considering all these requirements, we propose to use age, time, and gender-specific shared component models for predicting incidence rates in lethal cancers in which the expected corrrelation between incidence and mortality is high. Different models will be considered and their performance will be analyzed using brain cancer incidence and mortality data by gender and age-groups in 27 health units from Navarre and Basque Country (two Spanish regions) during the period 1998-2008. A fully Bayesian approach based on integrated nested Laplace approximations will be considered for model fitting and inference.

Keywords: Cancer prediction, disease-mapping, INLA, shared component models.

Acknowledgments This work has been supported by Proyecto Jóvenes investigadores de la UPNA (UPNA2018-11-INC) and partially funded by Fundación CAN-Obra Social Caixa- UNED Pamplona 2017.

Estimation of cancer incidence from mortality data: a validation study within a population-based cancer registry

Miguel Rodríguez-Barranco[1,2,3], *Daniel Redondo-Sánchez*[1,2,3], *Alberto Ameijide*[4],*Pablo Fernández-Navarro*[3,5], *Dafina Petrova*[1,2,3], *Maria José Sánchez*[1,2,3]

[1]Granada Cancer Registry, Andalusian School of Public Health.
[2]Biomedical Research Institute of Granada (ibs.Granada). University of Granada.
[3]CIBER of Epidemiology and Public Health (CIBERESP).
[4] Tarragona Cancer Registry, Fundation Society for Cancer Research and Prevention (FUNCA), Pere Virgili Health Research Institute (IISPV).
[5]Cancer and Environmental Epidemiology Unit, National Center for Epidemiology, Carlos III Institute of Health.

Background: Population-based cancer registries are required to calculate cancer incidence in a geographical area, and several methods have been developed to obtain estimations of cancer incidence in areas not covered by a cancer registry. However, an extended analysis of those methods in order to confirm their validity is still needed.

Objective: To evaluate the predictive performance of one of the most commonly used methods to derive cancer incidence rates from mortality and incidence-mortality ratio (IMR).

Methods: Using incident cases from all cancer sites (excluding non-melanoma skin cancer) during the period 1985-2008 from the Granada Cancer Registry, we compared incident cases estimated with the IMR method to observed cases diagnosed in 2004-2013 in Granada for total cancer and six cancer sites of interest for each sex. Using the previous 15-years mortality time series (1985-2010) and different functional forms of the IMR trend, we derived the expected yearly number of cancer cases for the period 2004-2013. We used GLMM including a polynomial function for calendar year of death and smoothing splines for age. To fit the models, we used a Bayesian framework based on MCMC. A goodness-of-fit indicator (GOF) was formulated to determine the best assumption of the IMR trend.

Results: 53096 cancer incidence cases and 43884 deaths due to cancer were included. The average relative deviation along the time series between the observed and predicted number of cancer cases for all cancer sites was 6% in men and 4% in women. Of the 12 cancer sites studied, 6 had deviations lower than 5%, and 8 lower than 10%. The constant assumption for the IMR trend provided the best GOF for rectum, lung, bladder and stomach cancers in men, and colorectal and corpus uteri in women. The linear assumption was better for colon in men, and lung and ovary cancers in women. The adjustment for breast and prostate cancer was worse than for the other cancer sites.

Conclusions: Overall, the IMR method showed good reliability for most cancer sites, except those with low lethality or sudden changes in incidence trends, in these situations, other methods are needed to get a suitable estimation. Funding: "Subprogram Cancer surveillance" of the CIBER of Epidemiology and Public Health (CIBERESP) and Andalusian Department of Health Research, Development and Innovation, project grant PI-0152/201

Keywords: Cancer incidence, validation, goodness-of-fit.

Estimation of reduced life expectancy among persons with a given disease and the importance of different causes of death: the Life Years Lost method, developed R functions, and example on mental disorders

Oleguer Plana-Ripoll[1], *John J. McGrath*[2], *Per K. Andersen*[3]

[1]pr@econ.au.dk, National Center for Register-based Research, Aarhus University .
[2]j.mcgrath@uq.edu.au, National Center for Register-based Research, Aarhus University.
[3]pka@biostat.ku.dk, Section of Biostatistics, University of Copenhagen.

Background: Life expectancy at a given age is a summary measure of mortality rates, which is estimated as the area under the survival curve, and represents the average number of years an individual at that age is expected to live if current age-specific mortality rates apply in the future. A complementary metric is the number of Life Years Lost, which is used to measure the reduction in life expectancy for a specific group of persons, e.g. those with a specific disease or condition. However, calculation of life expectancy among those with a disease is not straightforward for diseases that are not present at birth, and previous studies have considered a fixed age of onset of the disease, e.g. at age 15 years. A recently-introduced method takes into account the real age-of-onset distribution, which leads to more accurate estimates. In addition, the method allows to decompose differences in life expectancy according to different causes of death using a competing risks framework. The aim of this communication is to introduce the Life Years Lost method-including developed R functions - and to show a comprehensive analysis of mortality associated with mental disorders using Danish population-based registers.

The Life Years Lost method: Given a cohort of people with a disease or condition, the method uses age at disease diagnosis for each person as its starting point and estimates the expected residual lifetime at that age using age-specific mortality rates among the diseased. The number of excess Life Years Lost is estimated by comparing the expected residual lifetime among someone with the disease, with that of the reference population of same age. In order to obtain a single estimate of excess Life Years Lost instead of one for each person, it is possible to take an average of all the person-specific Life Years Lost. This estimate can be interpreted as the average number of Life Years Lost that patients with a given disease or condition experience from the time of diagnosis in excess to those experienced by a reference population of same age.

Example: We designed a population-based cohort study including all Danish residents between 1995 and 2015 (N = 6,619,045). For each specific mental disorder, Life Years Lost (LYLs) were estimated for all-cause mortality, and for each specific cause of death. We found that all types of mental disorders were associated with a shorter life expectancy (LYLs ranging from 5.2 years for organic disorders in women to 14.7 years for substance use disorders in men).

Keywords: Survival analysis, epidemiology, mental disorders.

Estimation of an energy poverty index at small area level based on the Barcelona health survey

Marc Marí-Dell'Olmo[1], *Carlos Vergara*[2], *Laura Oliveras*[3], *Miguel Ángel Martínez-Beneito*[4]

[1]mmari@aspb.cat, Servei de Qualitat i Intervenció Ambiental, Agència de Salut Pública de Barcelona.

[2]vergara_car@gva.es, Área de Desigualdades en Salud-Servicio de Estudios Estadísticos, Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana (FISABIO).

[3]loliveras@santpau.cat, Institut d'Investigació Biomèdica Sant Pau (IIB Sant Pau).

[4]martinez_mig@gva.es, Área de Desigualdades en Salud-Servicio de Estudios Estadísticos, Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana (FISABIO).

The Barcelona Health Survey (BHS) allows us to know and monitor the state of health and its determinants, as well as lifestyles and use of health services of the population of Barcelona. The sample size of the 2016 BHS is 4000 people and a subsample of 400 people is selected in each of the 10 districts of the city. This survey is designed to provide good representativeness at the district level and records are weighted by using sample weights that consider district, sex and age. The BHS has several energy poverty indicators and there is a growing interest in obtaining reliable estimates of these indicators in the 73 neighborhoods of the city, where there is under-representativeness of the sample. In addition, we want to obtain an energy poverty index, which synthesizes the information from all the indicators into a single variable. To this end, the Bayesian model "M-Model" has been imported from the field of disease mapping. This method offers distinct advantages over other techniques for dimensionality reduction, such as Principal Components Analysis (PCA) or Bayesian Factor Analysis. The "M-Model" allowed us to consider both the dependence between the different indicators analyzed and the existence of spatial dependence between their values. With this model it was possible to obtain smoothed energy poverty indicators and, to construct an energy poverty index with the first dimension of the PCA of its variance-covariance matrix. Finally, we could correct the possible under-representativeness of groups of individuals who are more vulnerable to energy poverty without the need to estimate sample weights, but considering their real distribution in the study population.

Keywords: Multivariate analysis, Bayesian model, spatial analysis.

Labour market participation patterns as predictors for sickness absence trajectories

Hernando-Rodriguez JC [1,2,3], *Serra-Saurina* [1,2,3],*Benavides FG* [1,2,3], *Ubalde-Lopez M* [1,2,3]

[1] Center for Research in Occupational Health (CiSAL), Pompeu Fabra University, Barcelona

[2] IMIM - Parc Salut Mar, Barcelona

[3] CIBER of Epidemiology and Public Health (CIBERESP), Madrid

Background: Changes in employment, working conditions and transitions between employment status characterise working life, which may influence an individual's future health-course. Prior research on sickness absence (SA) (i.e. absence from work due to a health problem) have focused on risk factors in the context of workplace based on one-time point, and just a few studies have studied the relationship between SA and labour market participation (LMP) from a longitudinal approach. This study aims to assess the association between prior LMP patterns and the course of future SA (2012-2014).

Methods: A cohort study based on a sample of 11,968 salaried workers affiliated with the Spanish Social Security system (62% women), living in Catalonia, with more than 15 accumulated days on SA per quarter during 2012-2014. We considered three working life cohorts according to the working life stage in 2002: early (18-25 years), middle (26-35 years) and late (36-45 years). In a first step, sequence analysis was used to construct individual LMP trajectories based on transitions among five employment states (employment, subsidised unemployment, unemployment with benefits, transition, no-affiliation with the Social Security) during 2002-2011. In a second step, optimal matching combined with cluster analysis was used to group similar sequences of weekly employment states. Latent class growth models were performed to identify SA trajectories based on accumulated days on SA quarterly during 2012-2014. Finally, crude and adjusted odds ratio (aOR) were estimated using multinomial logistic regression models. We consider employment conditions (the type of contract, working time, salary and occupational category), SA medical diagnosis and time employed the prior ten years as potential confounders.

Results: Overall we identified four SA trajectories: low stable (83%-88% of the workers), decreasing (5%-9%), increasing (5%-11%) and high stable (7%-16%) accumulated days on SA, for men and women. Likewise, six LMP patterns were obtained: stable employment (63%-81%), increasing employment (5%-22%), delayed employment (7%-8%), decreasing employment (4%-10%), varying employment (13%-14%), steeply decreasing employment (9%), and steeply labour market exit (7%-9%). LMP patterns were not significantly associated with SA trajectories. Exceptionally, among young men, the increasing employment pattern was related with a lower risk to increase accumulated days on SA over time (aOR: 0.21 [95% CI: 0.04-0.96]) compared to the low stable SA trajectory.

Conclusions: SA trajectories were not related to extended prior working life patterns. A closer working life exposure period to the SA follow-up might be considered when studying their relationship.

Keywords: Working-life transitions, life-course, sick leave.

Conferencias Invitadas:
Sesión SEB-EMR-Portugal

Construction of confidence intervals for the maximum of the Youden index and its corresponding optimal cutoff point

Leonidas E. Bantis[1], *Christos T. Nakas*[2], *Benjamin Reiser*[3]

[1]leobanits@gmail.com, Department of Biostatistics, University of Kansas Medical Center
[2]cnakas@gmail.com, Laboratory of Biometry, University of Thessaly
[3]reiser@stat.haifa.ac.il, Department of Statistics, University of Haifa

The maximum of the Youden Index is a frequently used summary measure of the ROC (Receiver Operating Characteristic) curve. In practice clinicians are in need of a cutoff point to determine whether intervention is required after establishing the utility of a continuous biomarker. The Youden index can serve both purposes as an overall index of a biomarker's accuracy, which also corresponds to an optimal (in terms of maximizing the Youden index) cutoff point that in turn can be utilized for decision-making. In this paper, we provide new methods for constructing confidence intervals for both the maximum of the Youden index and its corresponding cutoff point. We explore approaches based on the delta approximation under the normality assumption, as well as power transformations to normality and nonparametric kernel- and spline-based approaches. We compare our methods to existing techniques through simulations in terms of coverage and width. We then apply the proposed methods to serum-based markers of a prospective observational study involving diagnosis of late-onset sepsis in neonates.

Keywords: Box-Cox transformation, delta method, ROC curve.

Marginalized frailty-based illness-death model: application to the UK-Biobank survival data

Nir Keret[1], *Malka Gorfine*[2]

[1]keret.nir@gmail.com, Department of Statistics and OR, Tel-Aviv University
[2]malkago12@gmail.com, Department of Statistics and OR, Tel-Aviv University

The UK Biobank is a large-scale health resource comprising genetic, environmental and medical information on some 500,000 volunteer participants in the UK, recruited at ages 40–69 during the years 2006–2010. The project monitors the health and well-being its participants. In this work we demonstrate how these data can be used to estimate in a semi-parametric fashion the effects of genetic and environmental risk factors on the hazard functions of various diseases. To do so, we assume an illness- death model, which inherently is a semi-competing risks model, since the death can censor the disease, but not vice versa. We define three dependent random processes: time from healthy state to age-at-onset, time from healthy state to death, when the person is free of the disease, and time from healthy state to death when the person has been diagnosed with the disease. By using a shared-frailty (random-effects) approach to account for the dependency between these processes, the marginal hazard functions are estimated. In addition, the recruitment procedure introduces left-truncation to the data, since a person can only be observed in the sample if they have lived long enough to reach the recruitment age. An additional challenge arising from the recruitment procedure, is that information coming from both prevalent and incident cases must be aggregated, while these two sources of information should be treated differently in the model. Lastly, since we do not observe any deaths prior to the minimal recruitment age, 40, we cannot directly estimate the death hazard function from the data and must supplement our procedure with external data. In our work we extend the initially developed model so that it will account for left-truncation, and thereby overcome those challenges.

Keywords: Semi-competing risks, frailty model, left truncation.

Methods for checking the Markovian assumption in Multi-state models

Gustavo Soutinho[1], *Luís Meira-Machado*[2], *Pedro Oliveira*[3]
[1]gustavosoutinho@sapo.pt, EPIUnit, ICBADS, University of Porto;
[2]lmachado@math.uminho.pt, Centre of Molecular and Environmental Biology, University of Minho;
[3]pnoliveira@icbas.up.pt, EPIUnit, ICBADS, University of Porto

Multi-state models provide a relevant modeling framework to deal with complex longitudinal survival data in which individuals may experience more than one event of interest. The estimation of the transition probabilities is a central question in these models since they allow long term predictions of the process. These quantities are usually estimated by the Aalen-Johansen (AJ) estimator which assumes the process to be Markovian. The consistency of the AJ estimator is not guaranteed in situations in which the process is non-Markov leading in these cases to biased estimators. Recently Uña-Álvarez and Meira-Machado [2] introduced alternative estimators based on subsampling (also known as landmarking) (LM) which are consistent regardless the Markov assumption. The checking of the Markovian assumption is in this case a relevant issue for choosing the best estimator since the AJ estimator is the one with less variability but may be biased depending on the Markov assumption.

The Markov assumption claims that given the present state, the future evolution of the process is independent of the states previously visited and the transition times among them. Traditionally this assumption is checked by including covariates depending on the history through a proportional hazards model. Since the landmark methods of the transition probabilities are free of the Markov assumption, they can also be used to introduce such tests (at least in the scope of the progressive multi-state models) by measuring their discrepancy to Markovian (AJ) estimators. In this paper, we introduce two local tests for the Markov assumption: one based on the log-rank test, and a test using bootstrap built on the areas between the transition probabilities curves obtained from the two estimators. The validity and behavior of the proposed methods was evaluated through simulation studies. The applicability of the proposed method is illustrated using real data.

Keywords: Markov assumption, Multi-state models, transition probabilities.

References:

[1] Aalen, O.O. Nonparametric inference for a family of counting processes. Annals of Statistics, **1978**, *6*, 701–726

[2] de Uña-Álvarez, J.; Meira-Machado, L. Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study. Biometrics, **2015**, *71*, 141–150.

Survival analysis taking competing risks into account

Laetitia Teixeira[1], *Anabela Rodrigues*[2], *Denisa Mendonça*[3]

[1]lcteixeira@icbas.up.pt, Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto & CINTESIS-ICBAS, University of Porto & EpiUnit, University of Porto

[2]rodrigues.anabela2016@gmail.com, Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto & Department of Nephrology, Santo António Hospital

[3]dvmendon@icbas.up.pt, Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto & EpiUnit, University of Porto

Survival analysis is a statistical method widely used in medical literature that explores the time period from a certain point until the occurrence of the event of interest. In various areas of medical research, we are in the presence of multiple competing events. A competing risk is an event whose occurrence either precludes the occurrence of another event under examination or fundamentally alters the probability of occurrence of this other event. Classical statistical methods for the analysis of survival data assume that competing risks are absent. For this reason, it is crucial be aware of appropriate methods to account for competing risks when analyzing survival data.

Different approaches are available to explore survival data in the presence of competing risks. Cumulative indice function (CIF) can be used for statistical description; Cause-specific hazard model and Fine & Gray model could be used to identify potential predictive factors of survival. This work has as objective: i) to describe different approaches for survival analysis in the presence of competing risks and ii) to apply methods described in a real clinical data (peritoneal dialysis).

Chronic renal disease is recognized as a global public health problem, reflecting the increase in the number of patients in need of renal replacement therapy. In the last years we verify an increase in the use of peritoneal dialysis as renal replacement therapy. Motivated by this increase, it becomes mandatory to evaluate this treatment program. Patient survival and technique survival, having respectively as outcome the events "death" and "transfer to haemodialysis", stand out from the several indicators used to evaluate a peritoneal dialysis program. When evaluating these indicators in the context of peritoneal dialysis it is crucial to take competing risks into account.

Keywords: Survival analysis, competing risks.

Strategies for integrated analysis in Imaging Genetics

Natalia Vilor-Tejedor[1]

[1]natalia.vilortejedor@crg.eu, Centre for Genomic Regulation (CRG), the Barcelona Institute for Science and Technology, Barcelona, Spain. Barcelonabeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain.

Imaging Genetic (IG) studies integrate neuroimaging and genetic data from the same individual, offering the opportunity to deepen our understanding of the biological mechanisms behind neurodevelopmental domains and complex neurodegenerative diseases (NDs).

Genetic risk factors are of potential interest for research on preventive practices of NDs. Preventive practices make personalized medicine possible by targeting those risk factors, allowing an appropriate treatment or preventive strategies for people who are at increased genetic risk of developing specific NDs. However, for some NDs, such as Alzheimer's disease, genetic association findings do not explain their whole genetic architecture because their underlying complexity is not captured entirely by disease status. Hence, identifying neuroimaging-based features affected by these genes can increase our understanding of NDs and aid future functional studies.

Although the literature on IG studies has sustained exponential growth during these last years, the majority of studies have mainly analysed individual associations of candidate brain regions with genetic variants. However, this strategy was not designed to deal with the complexity of neurobiological mechanisms underlying NDs. Moreover, these studies require a large amount of samples to discover genetic effects that survive stringent multiple comparison corrections, and multidimensionality of the data represents a challenge for the standardization of modelling procedures.

We provide a systematic update of current methods and strategies used in IG studies, which may serve as an analytical framework for researchers working in the neurogenetics field. In addition, we present an overview of how these methodological approaches are applied for the integration of neuroimaging and genetic data. We hypothesize that relevant IG findings in relation to biological mechanisms (e.g., the identification of new target genes involved in brain structure and functioning) might assist with the design of personalized disease-modifying drugs and treatments, especially for NDs.

Keywords: Imaging Genetics, neurogenetics, multivariate modelling.

Conferencias Invitadas:
Sesión Jóvenes Invesigadores/as

Analysis of 2D Foot morphology by functional archetypal analysis

A. Alcacer[1], *I. Epifanio*[1], *M.V. Ibáñez*[1], *A. Simó*[1]

aalcacer@uji.es, epifanio@uji.es, mibanez@uji.es, simo@uji.es

[1]Departament de Matemàtiques, Universitat Jaume I, Spain

Abstract

Improving the fit in footwear is an important issue both for manufacturers and customers. For that reason, an anthropometric database of the adult Spanish population is analyzed. Shapes of feet are represented by 2D images. Archetype analysis is the appropriate statistical tool to describe the extreme patterns. We have extended Archetype analysis to functions with two arguments. We have applied this new methodology to the images of feet of women and men.

Keywords: Archetype analysis, shape analysis, functional data analysis.

1. Introduction

Knowledge of foot shape has a great relevance for the appropriate design of footwear. It is a very important issue for manufacturing shoes, since a proper fit is a key factor in the buying decision, besides improper footwear can cause foot pain and deformity, especially in women. Therefore, the objective is to identify the shapes that represent the fitting problems of the target population by means of archetypal shapes, which are extreme patterns. Then the shoe designer may adapt the design to the measurements of the extremes of a size. Archetype Analysis (AA) [1] is an unsupervised data mining technique that describes instances of a sample as a convex combination of certain number of elements called archetypes, which in turn, are convex combinations of the observations in the sample. This multivariate technique was extended to functional data with one argument ([3, 5]).

A 3D anthropometric study of the feet of Spanish adult population was carried out by the Instituto de Biomecánica de Valencia. We consider their footprints, i.e. 2D images, which can be seen as functions with two arguments and functional data techniques ([4]) are used. The purpose of this work is to extend functional archetype analysis (FAA) to 2D binary images, i.e. to functions with two arguments, and to apply it to the novel data set. Furthermore, FAA will help an image data set easier to understand, displaying and describing their features.

Section 2 describes our data. In Section 3 the methodology is introduced, and results are analyzed in Section 4. Finally, some conclusions are given in Section 5.

2. Data

Footprints have been extracted from an anthropometric database of 775 3D right foot scans representing Spanish adult female and male population, 382 correspond to women and 393 to men. Data was collected in different regions across Spain at shoe shops and workplaces using an INFOOT laser scanner. The binary images have been centered and scaled in order to remove the effects of translations and changes of scale and to consider only the shape, as explained by [2].

3. Methodology

3.1 AA for (standard) multivariate data

Let $\mathbf{X}$ be an $n \times m$ matrix with n cases and m variables. The objective of AA is to find k archetypes, i.e a $k \times m$ matrix $\mathbf{Z}$, in such a way that $\mathbf{x}_i$ is approximated by a mixture of $\mathbf{z}_j$'s (archetypes):

$$\sum_{j=1}^{k} \alpha_{ij} \mathbf{z}_j, \tag{1}$$

with the mixture coefficients contained in the $n \times p$ matrix α.

Additionally, $\mathbf{z}_j$'s is expressed as a mixture of the data through the mixture coefficients found in the $k \times n$ matrix β:

$$\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl} \mathbf{x}_l. \tag{2}$$

To obtain the archetypes, AA computes two matrices α and β that minimize the following residual sum of squares (RSS): $\sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \mathbf{z}_j\|^2$ = $\sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathbf{x}_l\|^2$, under the constraints 1) $\displaystyle\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and 2) $\displaystyle\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ for $j = 1, \ldots, k$.

3.2 AA for functional data

In the functional context, the values of the m variables in the standard multivariate context are replaced by function values with a continuous index t. Similarly, summations are replaced by integration to define the inner product. See [3] for details about extension of AA to functional data. Here, we extend the work by [3] to functional data with two arguments.

Let $f_i(s,t)$ be a function defined in $[a,b] \times [c,d]$. Its squared norm is $\|f_i\|^2$ = $\int_c^d \int_a^b f_i(s,t)^2 dsdt$. In the basis approach, each function f_i is expressed as a linear combination of known basis functions B_h with h = 1, ..., m: $f_i(s,t)$ = $\sum_{h=1}^{m} b_i^h B_h(s,t)$ = $\mathbf{b}_i'\mathbf{B}$, where $'$ stands for transpose and $\mathbf{b_i}$ indicates the vector of length m of the coefficients and $\mathbf{B}$ the functional vector whose elements are the basis functions. In FAA with two arguments the objective functions is: RSS = $\displaystyle\sum_{i=1}^{n} \|f_i - \sum_{j=1}^{k} \alpha_{ij} z_j\|^2$ = $\sum_{i=1}^{n} \|f_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} f_l\|^2$ = $\sum_{i=1}^{n} \|\mathbf{b'}_i\mathbf{B} - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathbf{b'}_l\mathbf{B}\|^\mathbf{2}$ = $\sum_{i=1}^{n} \|(\mathbf{b'}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathbf{b'}_l)\mathbf{B}\|^\mathbf{2}$ = $\sum_{i=1}^{n} \|\mathbf{a'}_i\mathbf{B}\|^2 = \sum_{i=1}^{n} < \mathbf{a'}_i\mathbf{B}, \mathbf{a'}_i\mathbf{B} >$ = $\sum_{i=1}^{n} \mathbf{a'}_i\mathbf{W}\mathbf{a}_i$, with the corresponding constraints for α and β; and where $\mathbf{a'}_i$ = $\mathbf{b'}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathbf{b'}_l$ and $\mathbf{W}$ is the order m symmetric matrix with elements w_{m_1,m_2} = $\int_c^d \int_a^b B_{m_1} B_{m_2} dsdt$, i.e. the matrix containing the inner products of the pairs of basis functions. In the case of an orthonormal basis, $\mathbf{W}$ is the order m identity matrix, and FAA is reduced to AA of the basis coefficients. But, in other cases, we may have to resort to numerical integration to evaluate $\mathbf{W}$, but once $\mathbf{W}$ is computed, no more numerical integrations are necessary.

4. Results

Images have been expressed in the 2D Discrete Cosine Transform, which is an orthonormal base. We have only considered the 0.62% of the first coefficients, since details of feet are kept with these number of coefficients. We have applied FAA for the groups of women and men, separately. In both cases, we have represented the screeplot, with the number of archetypes versus the respective RSS, and we have found an elbow at k = 2, in both cases. Figure 1 shows the archetypes (A_1 and A_2) for women and men. The intersection between both archetypes is displayed in white, while the set difference A_1 A_2 is displayed in red, and the set difference A_2 A_1 is displayed in blue.

Results for women and men are very similar, so they are commented both together. In both cases, the difference in shapes of A_1 and A_2 are in the lateral zones in front of the top and bottom zones of the foot image. We have also displayed the α values of A_1 versus the foot length for women and men, and points are uniformly distributed. Therefore, the different shapes are found in all the sizes.

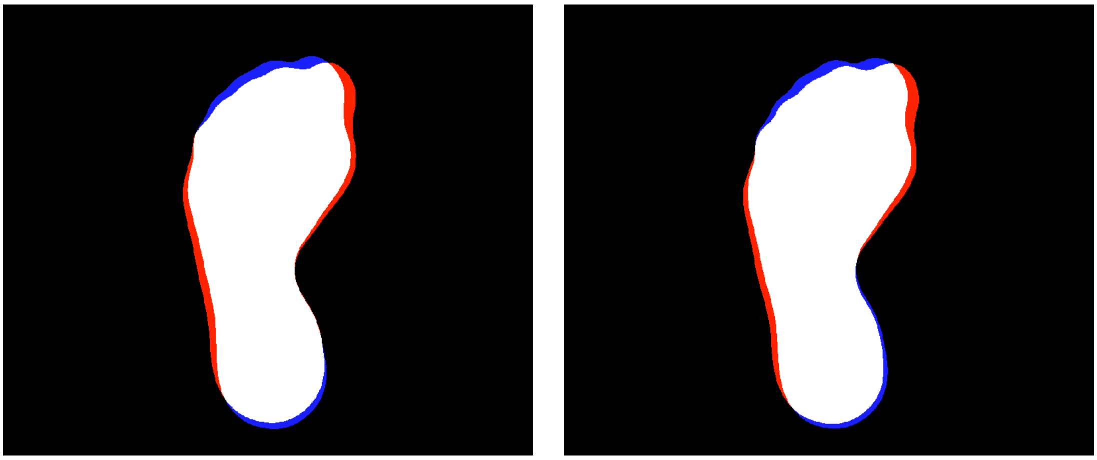


Figure 1: Archetypes for women (left-handed) and men (right-handed). See text for details.

5. Conclusions

AA has been extended to functions with two arguments, and we have applied it to a novel data set of foot images. Knowing the extreme shapes can help shoe designers adjust their designs to a larger number of the population and be aware of the characteristics of the users that will not be comfortable to use them, whether to consider a line of special sizes or modify any shoe feature to cover more customers.

As future work, we can extend AA to functions with three arguments in order to analyze 3D foot shapes.

Acknowledgments

This work is supported by the grant DPI2017-87333-R from the Spanish Ministry of Science, Innovation and Universities (AEI/FEDER, EU).

Bibliography

[1] Cutler, A., Breiman, L.: Archetypal Analysis. Technometrics **36**(4), 338–347 (1994)

[2] Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis: With Applications in R, 2nd edn. Wiley (2016)

[3] Epifanio, I.: Functional archetype and archetypoid analysis. Computational Statistics & Data Analysis **104**, 24 – 34 (2016)

[4] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer (2005)

[5] Vinué, G., Epifanio, I.: Archetypoid analysis for sports analytics. Data Mining and Knowledge Discovery **31**(6), 1643–1677 (2017)

[6] Vinué, G., Epifanio, I., Alemany, S.: Archetypoids: A new approach to define representative archetypal data. Computational Statistics & Data Analysis **87**, 102 – 115 (2015)

A nonparametric two-sample class of statistics for time-to-event and binary outcomes

Marta Bofill Roig[1] *and Guadalupe Gómez Melis*[1]

[1]marta.bofill.roig@upc.edu. Universitat Politècnica de Catalunya, Barcelona Graduate School of Mathematics

Abstract

We propose a novel class of tests to jointly evaluate the efficacy on binary and survival endpoints. The proposed class of statistics tests the equality of survival functions and the equality of proportions in a two-arm controlled trial. The proposed statistics are fully non-parametric and do not need the proportional hazards assumption for the survival endpoint.

Keywords: Clinical trial, failure time data, Pepe-Fleming statistic.

1. Introduction

The design and analysis of randomized controlled trials comparing the effectiveness of two or more interventions involves several phases of clinical research. The drug development process in oncology trials begins with a phase I trial to identify the appropriate drug dose and pharmacological characteristics. This is followed by a phase II trial that tests the safety and efficacy of a fixed dose of the new drug. Common efficacy endpoints for phase II trials are binary endpoints based on the tumor size such as objective response. If the drug has shown promising results in phase II trials, then the safety and efficacy are tested in a double-blinded randomized phase III trial with a much larger sample size. The efficacy endpoints of a phase III trial are often time-to-event endpoints being overall survival the gold standard.

The emergence of cancer immunotherapies has opened new statistical challenges due to two different reasons. First, these novel therapies induce a delayed start of treatment effect which causes that the proportional hazards assumption does not hold. The non-proportionality of hazards implies that the hazard ratio has not a clear interpretation and that the commonly used logrank test may result in a loss of statistical power. Second, since traditional endpoints may not capture the clinical benefit of cancer immunotherapies, this leads to the need to refine clinical trial endpoints. Novel endpoints that capture both tumor response and improved survival may be appropriate for a better characterization of the clinical response.

When multiple outcomes are collected and measured in different scales –such as continuous and binary responses–, the usual modeling strategy is to consider each outcome separately in a univariate framework. This strategy, however, ignores the correlation between the outcomes and thus could give a less efficient design. Likelihood-based multivariate approaches have been proposed to model mixed outcomes accounting for the association between them. Parametric mixture models have been suggested to bind the binary tumor response in phase II to the survival endpoint in phase III in a Bayesian framework. However, to the best of our knowledge, a joint test for two-sample comparison on survival and binary outcomes has not been proposed. In this paper, we propose a novel class of statistics to jointly evaluate the efficacy on binary and survival endpoints that could be used in a seamless phase II/III design.

2. A general class of survival and binary statistic tests

Consider a two-arm randomized controlled trial with follow-up τ_s and let ε_s and ε_b be events of interest. Suppose that the corresponding endpoints for these events are the time until ε_s, denoted by T, and the binary outcome of having had ε_b at certain pre-specified time-point τ_b ($\tau_b < \tau_s$), denoted by X. For instance, in oncology trials, these endpoints could be overall survival and objective response, respectively.

Assume that the goal of the trial is to determine whether there is an effect on at least one of these two endpoints, so that we aim to test the null hypothesis of non-effect on the survival endpoint neither on the binary endpoint against the alternative hypothesis of an effect on at least one of these endpoints. The hypothesis problem can then be formalized as:

$$\begin{cases} \mathrm{H}_0: & p^{(0)} = p^{(1)} \ \text{ and } \ S^{(0)}(t) = S^{(1)}(t) \ \text{ for all } \ t \in [0, \tau_s] \\ \mathrm{H}_1: & p^{(0)} < p^{(1)} \ \text{ or } \ S^{(0)}(t) < S^{(1)}(t) \ \text{ for } \ t \in [0, \tau_s] \end{cases} \tag{1}$$

where $S^{(i)}$ is the survival function of T and $p^{(i)}$ is the probability of $X = 1$ for the i-th group ($i = 0, 1$). Note that this null hypothesis is an intersection of two sub-hypotheses $\mathrm{H}_0 : \mathrm{H}_{b,0} \cap \mathrm{H}_{s,0}$ given by $\mathrm{H}_{b,0} : p^{(0)} = p^{(1)}$ and $\mathrm{H}_{s,0} : S^{(0)} = S^{(1)}$.

We propose a class of statistics for testing (1) using a linear combination of statistics for each one of the sub-hypotheses $\mathrm{H}_{b,0}$ and $\mathrm{H}_{s,0}$. Let $Z_{k,n}$ be the test statistic for the sub-hypothesis $\mathrm{H}_{k,0}$ ($k = s, b$) and assume that, under the null hypothesis $\mathrm{H}_{k,0}$, $Z_{k,n}$ follows the standard normal distribution. The class of statistics –hereafter called $\mathcal{L}$-class– is defined by:

$$Z_{BS,n} \;=\; \frac{1}{\sqrt{2 + 2 \cdot \mathrm{Cov}(Z_{b,n}, Z_{s,n})}} \cdot (Z_{b,n} + Z_{s,n}) \tag{2}$$

Note that under H_0, $Z_{BS,n}$ follows the standard normal distribution. The statistics of this class reject H_0 for large values of the sum of the respective statistics for the sub-hypotheses $\mathrm{H}_{b,0}$ and $\mathrm{H}_{s,0}$. In what follows we present the considered statistics $Z_{k,n}$ for this work and the properties of the subsequent $Z_{BS,n}$.

Choice of the statistic $Z_{b,n}$: We consider the score test for testing the equality of proportions, $\mathrm{H}_{b,0}$, defined by:

$$Z_{b,n} \;=\; \frac{\sqrt{\frac{n^{(0)} n^{(1)}}{n^{(0)}+n^{(1)}}} \cdot \left(\hat{p}^{(0)} - \hat{p}^{(1)}\right)}{\sqrt{\frac{\left(\hat{p}^{(0)}+\hat{p}^{(1)}\right)\left(1-\hat{p}^{(0)}+\hat{p}^{(1)}\right)}{4}}} \tag{3}$$

where $\hat{p}^{(i)}$ is the estimated proportion of events ε_b and $n^{(i)}$ is the sample size for the i-th group.

Choice of the statistic $Z_{s,n}$: For testing the equality of two survival functions, $\mathrm{H}_{s,0}$, we consider the standardized weighted Kaplan-Meier statistic [1] given by:

$$Z_{s,n} = \frac{\sqrt{\frac{n^{(0)} n^{(1)}}{n^{(0)}+n^{(1)}}} \int_0^{\tau_s} \hat{w}(t) \left(\hat{S}^{(1)}(t) - \hat{S}^{(0)}(t)\right) dt}{\sqrt{- \int_0^{\tau_s} \frac{\left(\int_t^{\tau_s} \hat{w}(u) \hat{S}(u) du\right)^2}{\hat{S}(t) \hat{S}^{-}(t)} \cdot \frac{n^{(0)} \hat{C}^{(0)}(t) + n^{(1)} \hat{C}^{(1)}(t)}{\left(n^{(0)}+n^{(1)}\right) \hat{C}^{(0)}(t) \hat{C}^{(1)}(t)} d\hat{S}(t)}} \tag{4}$$

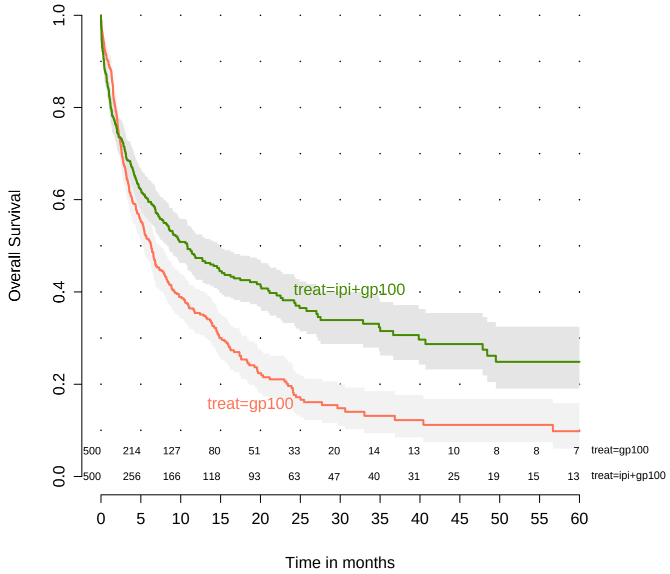
where $\hat{S}^{(i)}(t)$ and $\hat{C}^{(i)}(t)$ denote Kaplan-Meier estimators of the time-to-event and censoring survival functions of group i, respectively; $\hat{S}(t)$ is the Kaplan-Meier estimator calculated from the pooled sample. The term $\hat{w}(t)$ is a possibly random function that takes small values at the end of the follow-up period in the presence of heavy censoring. Note that $Z_{s,n}$ is based on the integrated weighted differences in Kaplan-Meier estimators and that, when $\hat{w}(t) = 1$, can be interpreted as the standardized difference of areas under the survival curves from $t = 0$ to $t = \tau_s$.

The resulting statistic of the $\mathcal{L}$-class constructed by using (3) and (4) does not require the proportional hazards assumption and is sensitive against the general alternative of stochastic ordering. Moreover, it can be shown that, under H_0, if $Z_{b,n}, Z_{s,n}$ are in the same direction, $\mathrm{Cov}(Z_{b,n}, Z_{s,n})$ is bounded between 0 and 1, and consequently $\frac{Z_{b,n}+Z_{s,n}}{2} \leq Z_{BS,n} \leq \frac{Z_{b,n}+Z_{s,n}}{\sqrt{2}}$.

3. Case Study: Vaccine therapy for metastatic melanoma

Melanoma has been considered a good target for immunotherapy and its treatment has been a key goal in recent years. Here we consider a randomized, double-blind, phase III trial whose primary objective was to determine the safety and efficacy of the combination of a melanoma immunotherapy (gp100) together with an antibody vaccine (ipilimumab) in patients with previously treated metastatic melanoma [2]. Despite the original endpoint was objective response rate at week 12, it was amended to overall survival and then considered secondary endpoint. A total of 676 patients were randomly assigned to receive ipilimumab plus gp100, ipilimumab alone, or gp100 alone. The study was designed to have at least 90% power to detect a difference in overall survival between the ipilimumab-plus-gp100 and gp100-alone groups at a two-sided α level of 0.05, using a log-rank test. Cox proportional-hazards models were used to estimate hazard ratios and to test their significance. The results showed that ipilimumab with gp100 improved overall survival as compared with gp100 alone in patients with metastatic melanoma. However, the treatment had a delayed effect and an overlap between the Kaplan-Meier curves was observed during the first six months. Hence, the proportional hazards assumption appeared to be no longer valid, and a different approach would had been advisable.

In order to illustrate our proposal, we consider the comparison between the ipilimumab-plus-gp100 and gp100-alone groups based on the overall survival and objective response rate as a co-primary endpoints of the study. For this purpose, we have simulated a dataset based on the information found on the full original protocol, the list of amendments and the statistical analysis plan, as well as, the main results and supplementary material. The simulation parameters were adjusted in order to mimic the delayed pattern observed in the real study. We have considered a copula-based framework to jointly simulate bivariate binary and time-to-event data. We have used a Clayton copula with parameter 0.5 and specified the marginal distributions so that: the time until overall survival follows a Weibull distribution with shape and scale parameters $a = 0.5$ and $b = 1$, an effect in the treatment group of $\mathrm{HR} = 0.7$ starting from the 6-th month; while the probability of having the objective response is $p^{(0)} = 0.1$ and $p^{(1)} = 0.05$ in the control and treatment group, respectively. The censoring distributions between groups were assumed equal and exponential with parameter 1; and the sample sizes were balanced and equal to 500. We have considered the weight function proposed in [1], that is, the pooled estimator for the censoring survival function $\hat{w}(t) = \hat{C}(t)$. The results are summarized in Figure and Table 1. Note that the combined statistic is bounded below by the value of the statistic itself calculated taking the covariance equals to 1, that is, $Z_{BS,n} \geq (Z_{b,n} + Z_{s,s})/2$. Computing the average of the individual statistic tests, we obtain that the lower value of $Z_{BS,n}$ is larger than the critical value z_α. We can then conclude that $Z_{BS,n} > z_\alpha$ and hence reject the null hypothesis H_0 of equal survival functions for the overall survival and equal proportions for the objective response.



Estimated parameters
Difference proportions: $\hat{p}^{(1)} - \hat{p}^{(0)} = 0.028$ Difference mean survival times: $\int_0^{\tau_s} \left(\hat{S}^{(1)}(t) - \hat{S}^{(0)}(t) \right) dt = 5.09$
Statistics
$Z_{b,n} = 2.60$ $Z_{s,n} = 4.79$
$Z_{BS,n} \geq \frac{Z_{b,n} + Z_{s,n}}{2} = 3.70$

Figure 1: Kaplan-Meier curves for ipilimumab-plus-gp100 and gp100-alone groups. Table 1: Estimated difference mean survival times and proportions, and the statistics described in (2), (3), and (4).

4. Results and Future Research

We have proposed a new class of statistics for testing the differences in two groups based on a binary and time-to-event endpoints. The statistics of the $\mathcal{L}$-class are sensitive to stochastic ordering alternatives for the survival endpoint and risk differences for the binary endpoint. Our ongoing research covers the derivation of the covariance and the study of the large sample properties of the $\mathcal{L}$-class. This work has been restricted to those cases in which the occurrence of the survival event or censorship does not prevent to assess the binary endpoint response. We are currently working on a more general censoring scheme where the binary endpoint could be censored.

Bibliography

[1] Pepe MS, Fleming TR. (1989). *Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data*. Biometrics, 45(2):497–507.

[2] Hodi, FS, et al. (2010). *Improved Survival with Ipilimumab in Patients with Metastatic Melanoma*. The New England Journal of Medicine, 363(8), 711-723.

Addressing alternative approaches for spatial modeling of herbicide retention in soil

Giannini Kurina F.[1], *S. Hang*[2], *A. Rampoldi*[1,2], *M. Cordoba*[1,2], *E. R. Macchiavelli*[3], *M. Balzarini*[1,2]

[1] francagianninikurina@gmail.com, Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina.

[2] Facultad de Ciencias Agropecuarias Universidad Nacional de Córdoba.

[3]Universidad de Puerto Rico, Mayagüez.

Abstract

Glyphosate retention coefficient (Kd) is modeled as function of soil variables, from a regional sampling, using: Ordinary and Partial Least Square regression, Random Forest, Generalized Boosted regression (GB), and Bayesian modelling with INLA; all regressions were fitted using spatial constraint on residuals. INLA produced the best fit, but GB the best spatial prediction.

Keywords: Predictive model, spatial data.

1. Introduction

Diffuse pollution derived from the use of plant protection products is a problem associated with agricultural intensification [1]. It requires deeper studies to adapt the available knowledge and generate useful technologies for decision making. The use of herbicides carries environmental risks that depends on the pesticide molecule and environmental characteristics (soil, climate, and management). Models to evaluate environmental hazards require, as inputs, variables that synthesize the interaction between herbicide and soil. An interaction that regulates pesticide behaviour in soil is retention, which is parameterized by the adsorption coefficient (Kd) [2]. It is a continuous variable that expresses the relationship between both, the amount of herbicide retained and the amount that remains in soil solution. Low values of Kd are often related to losses of leaching and runoff, while the potential losses by soil erosion are associated with high Kd values. Several soil variables may be conditioning the retention of the herbicide molecule in a site [3]. These variables are often spatially structured whereby modeling herbicide retention, through Kd coefficient, requires accounting for the underlying spatial variability. In general terms, a linear model for spatial data contains a deterministic component and a random one, which is explained by the spatial autocorrelation process and a net residual term. Models for spatial prediction can be generated from different strategies that allow fitting both, the deterministic component and the random one. In this work, approaches of different nature from which it is possible to adjust a regression model, including the spatial autocorrelation in the residual term, are addressed. We compare the results of three approaches to fit predictive regression models for spatial data: the frequentist [4] , the Bayesian [5], [6] and the Boosting-based[7].

2. Materials and Methods

A soil survey was conducted in Córdoba province, Argentina (29^o to 35^oS, and 61^o to 65^oW) that collected samples from the upper 15 cm of soil using a regular 40×40 km grid (90 sites)[3]. For each soil sample, the following variables were obtained: pH, total nitrogen, total organic carbon, Na, K, Ca, Mg, Zn, Mn, Cu, cation exchange capacity, the percentage of sand, silt and clay, water holding capacity and aluminium and iron oxides. The glyphosate Kd was determined in lab according to the bach-equilibrium technique for the preparation of soil suspensions. The concentration of herbicide in the soil solution under equilibrium (Ceq) was quantified by high pressure liquid chromatography (HPLC) according to Marek and Koskinen (2014). The adsorbed concentration (Cad) was calculated as the difference between the initial concentration and the concentration at equilibrium in the solution. The Kd was obtained as Cad/Ceq. The Kd values were transformed to the log scale because of the skewness distribution. To help the selection of soil variables that best explain Kd variability a regression tree improved by resampling (Boosting Regression Tree) [8], was used. The R package gbm, with the functions gbm.step and gbm.simplify, was implemented to select the subset of variables that minimize deviance. To fit predictive model for Kd, we assessed the following strategies: Multiple Linear Regression (ML) with spatially correlated errors through an spatial exponential function [9]; Random Forest Regression (RF), and Generalized Boosted Regression (GB), with spatially correlated residuals according with the methodology proposed by proposed by [7]. In the implementation an exponential model was fitted to the empirical semivariogram of the residuals of both machine learning algorithm. The same procedure to account for spatiality was implemented on the residuals obtained after fitting a Partial Least Square regression model (PLS). Additionally, a Bayesian model approached with Integrated Nested Laplace Approximation (INLA) [6] was fitted on the same data. To account for spatial correlation during the Bayesian modeling the Matern function solved by spatial partial differential equation (SPDE) [10], on the spatially structured random component was used. The complete R code to fit predictive models presented here is available at https://github.com/francagiannini. For all models, the mean square error (MSE) was obtained from the differences between observed and predicted value. In the Bayesian framework, it was calculated from the difference between the observed value and the mean of the posterior distribution for a missing data. The predictive ability of all compared methods was assessed using the leave-one-out cross-validation method to produce a global error measurement. The Mean Squared Prediction Error (MSPE) was obtained averaging the differences between the observed Kd value with the predicted one at each site. Additionally, a punctual prediction error, expressed as percentage of the Kd at each site, was calculated for all methods. It was referred as Site-Specific-Error (SSE) and categorized as smaller than 20%, between 20% and 40%, and greater than 40% of the site Kd, to visually interpret the goodness of prediction. The spatial patterns of the predictions (SEE mapping) of all methods were examined for their validity.

3. Results and Discussion

Predictive modeling [11] is the process that provides a mathematical tool (model) to predict an output from a convenient selected set of data. It demands the full understanding of the undergoing data generating process, model fitting, and its validation. In this study, we compare the results of three approaches to fit predictive regression models for spatial regional data. The soil variables of greater relevance to explain the variability of Kd?s in Cordoba, as indicated by the BRT algorithm, were aluminum oxides, pH, sand percentage and clay percentage. Consequently, all predictive models were fitted using those soil properties as explanatory variables. The Bayesian model with INLA produced the best fit (MSE=3.9% of the average Kd). However, in the cross-validation process to measure predictive ability, the Bayesian model was overcome by the PLS regression with spatial correlated residuals. The lowest MSPE was 18.9% of the average Kd mean (Table 1). The relation between MSE (a measure of goodness of fit), and MSPE (a measure of predictive ability) suggest that the Bayesian modeling with INLA and the SPDE, can overfit data. Thus, the PLS regression, accounting for spatiality, was a good approach in term of global error measurements. This result is probably explained by the collinearity among input variables (being the highest correlation coefficient equal to -0.75 between sand and clay fractions, p<0.001). The asymmetrical nature of Kd distribution and the multicollinearity among inputs made Boosted-base algorithms competitive. Such machine learning methods had been reported as more robust under these conditions [12]. Boosting-based methods, like RF and GB, have shown their superior performance in various disciplines, but they are commonly used with non-spatial data. Some uses of these methods with spatial[13], requires accounting the spatial structure through distance measurements which are incorporated as explanatory variables in the model [14]. However, as implemented here, spatiality was modeled through an autocorrelation spatial process on the residuals [7] which is easier to implement from a computational perspective.

Model	MSE[%]	MSPE[%]
ML	25.2	27.3
PLS	10.2	18.9
RF	11.9	19.9
GB	11.2	19.5
INFLA	3.9	25.4

Table 1: Goodness of fit (MSE) and average predictive ability (MSPE) of alternative regression models for glyphosate soil adsorption coefficient as a function of four soil variables. Ordinary (ML) and Partial Least Square regression (PLS), Random Forest (RF) and Generalized Boosted regression (GB); Bayesian modeling (INLA).

A deeper observation of the SSE showed that most Kd were well predicted in most of the sites (Figure 1) and high SSE where closely located (Northwest of Cordoba). In these sites, SSE was far superior to 40% and they consequently raised the global error measurement. However, it is important to highlight that the sites with high SSE (red points Figure 1) had extremely low Kd values and even with high prediction error, they are classified as low Kd sites. Thus, these SSE did not lead to misunderstanding of glyphosate retention.

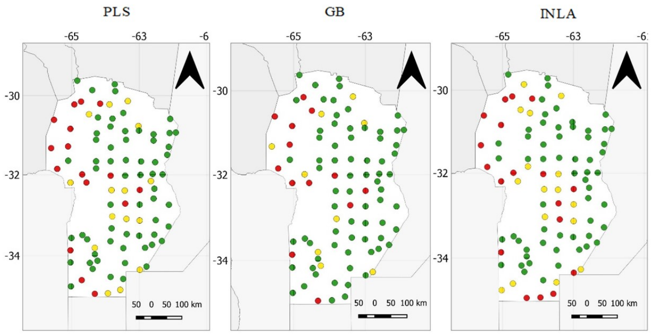


Figure 1: Site specific errors (SSE) for a model of glyphosate soil adsorption coefficient as a function of aluminum oxides, pH, clay, and sand at the site. SSE was categorized as smaller than 20% (green), between 20% and 40% (yellow), and greater than 40% (red) of the site Kd.

In the Fig. 1 we show that the GB model improved the SSE pattern in spatial predictions with respect to both PLS and INLA. GB was the procedure that presented the biggest proportion of prediction errors smaller than 20% of the site mean. Then GB regressions, with spatially correlated errors, can produce competitive results with respect to the frequentist and the Bayesian approaches. The GB advantage is that it requires much less statistical assumptions, and it is easier to automate. However, a better understanding of the process may require models that explicitly show the impact of each input variable like, the produced with INLA which best fitted the observed data. Further work on modeling the Glyphosate Kd distribution in

Córdoba may demand modeling strategies which contemplate mixture of distribution because the Kd values in the Northwest it may come from a different biological process than the others.

4. Conclusions

Visual examination of site-specific prediction errors proved to be an essential tool in assessing the spatial predictions. This study has simultaneously compared alternative methods for spatial interpolation of an environmental property (Glyphosate adsorption coefficient). Results confirm the effectiveness of Bayesian modeling with INLA to obtain good fitting, and the high predictive ability of GB regression in the context of models with several covariables and a spatial autocorrelation process underlying in the random component.

Bibliography

[1] Holland J. M. (2004). *The environmental consequences of adopting conservation tillage in Europe: reviewing the evidence*, Agric. Ecosyst. Environ.,vol.103, no.1, 1-25.

[2] Calvet R. (2005). *Les pesticides dans le sol: conséquences environnementales*. France Agri Editions.

[3] Hang S. , Negro G., Becerra A., and Rampoldi A. E. (2015). *Suelos de Córdoba: Variabilidad de las propiedades del horizonte superficial*. Córdoba, Argentina: Jorge Omar Editorial.

[4] Webster R. and Oliver M. A. (2007). *Geostatistics for environmental scientists*, vol. 1, no. 2. John Wiley & Sons.

[5] Wang X., Ryan Y. Y., and Faraway J. J. (2018). *Bayesian Regression Modeling with INLA*. Chapman and Hall/CRC.

[6] Blangiardo M. and Cameletti M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

[7] Li J., Heap A. D., Potter A., and Daniell J. J. (2011). *Application of machine learning methods to spatial interpolation of environmental variables*, Environ. Model. Softw., vol. 26, no.12,1647-1659.

[8] Elith J., Graham C. H., Anderson R. P., Dudik M., Ferrier S., Guisan A., Hijmans R. J., Huettmann F., Leathwick J. R., and Lehmann A. (2006). *Novel methods improve prediction of species distributions from occurrence data*, Ecography (Cop.)., vol. 29, no. 2, 129-151.

[9] Pinheiro J., Bates D., DebRoy S. , and Sarkar D.(2010). *R package*, Version 3.

[10] Krainski E. T. and Lindgren F. (2013). *The R-INLA tutorial: SPDE models Warning: work in progress... Suggestions are welcome to elias@ r-inla. org*.

[11] Kuhn M. and Johnson K. (2013). *Applied predictive modeling*, vol. 26. Springer.

[12] Duffy N. and Helmbold D. (2002). *Boosting methods for regression*, Mach. Learn.,vol.47,no. 2-3, 153-200.

[13] Kanevski M., Timonin V., and Pozdnukhov A. (2009). *Machine learning for spatial environmental data: theory, applications, and software*. EPFL press.

[14] Hengl T., Nussbaum M., Wright M. N., Heuvelink G. B. M., and Gräler B. (2018). *Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables*, PeerJ, vol. 6, p. e5518.

ORI. An Order Restricted Inference framework to analyse chronobiological rhythms

Yolanda Larriba[1], *Cristina Rueda*[1], *Miguel A. Fernández* [1] *and Shyamal D. Peddada* [2]

[1]yolanda.larriba@uva.es. Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Spain

[2]Department of Biostatistics, Public School of Health, University of Pittsburgh, USA

Abstract

This work is motivated by the problem of discovering temporal rhythmic patterns in oscillatory systems. We first develop a flexible mathematical formulation of rhythmicity, based on order restrictions. Then, we address commonly encountered problems in chronobiology. Specifically, we solve the problem of detecting rhythmic signals and those of estimating sample times. In both cases, our methodology yields better results than existing ones in literature.

Keywords: Order Restricted Inference, rhythmicity detection, oscillatory systems.

1. Introduction

Blood pressure, body temperature or circadian gene-expressions are just a few of the biological phenomena exhibiting rhythmic processes in nature. Such processes display periodic up-down-up patterns, or rhythms, along periods of time, usually of 24 hours length. The study of these temporal rhythms and how they change under different conditions is called chronobiology. For the last two decades, research on chronobiology has had a marked effect on preventing cardiovascular disorders like hypertension, on improving the effectiveness of cancer treatments or on detecting gene-expressions linked to diseases. These and other implications in health motivate a raised interest in chronobiology, in order to identify and/or characterize those rhythmic processes. From a statistical point of view, the modelling of chronobiological rhythms is a challenge as observed data usually present low sampling density along a short number of periods. Moreover, rhythm-patterns can adopt a wide range of rhythmic shapes which are not always well fitted using standard parametric models, as those based on cosine functions, because they are too rigid for rhythms derived from biological systems [1]. In addition, data derived from biological systems are usually related to highly noisy experiments, as in the case of circadian gene-expressions [2].

The main statistical problem to be solved in this context, and the most important one in practice, is that of identifying which patterns on observed data correspond to rhythmic processes and which do not. A wide variety of procedures to detect rhythmicity are available in literature. One of the most popular among biologists is JTK_Cycle (JTK) [3], based on the Jonckheere-Terpstra test and the Kendall's tau correlation. More recently, [1] presented a novel algorithm based on ORI to detect and classify two-period rhythms derived from the circadian system. It is important to note that the aforementioned algorithms, as well as most processes in literature to analyse rhythmicity, consider that the timing of the samples (c.f. the time of the day at which samples were taken in the case of the circadian gene-expressions, or the peak time for the genes participating in the cell-cycle) is a priori known. Yet, there exist experiments where the timing of the samples is unrecorded, since it may be impractical or dangerous, as in the case of human organ biopsies. When this happens, the temporal order of samples must be previously estimated before addressing any other question related to rhythmicity. To our knowledge, this relevant rhythmicity question has been barely dealt in literature. Recently, [4] proposed CYCLOPS, a complex approach based on machine learning and neural networks.

In addition to these two major problems (rhythmicity detection and timing estimation), other minor rhythmicity-related problems are raised due to its implications on health. The estimation of the time at which gene-expressions reach their peaks or the rhythm-pattern comparisons along different species are just a few of the additional topics analysed in recent biomedical studies.

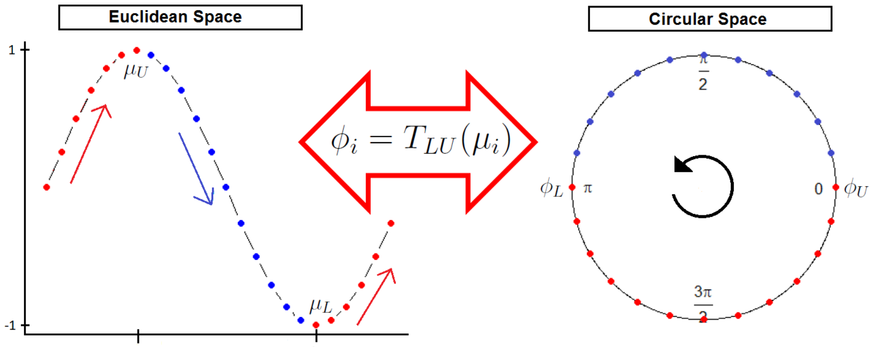
Solutions given until now are specific to the experiment discussed in each paper existing a lack of uniformity in practical procedures. It is highly desirable to establish a general statistical framework to formulate and solve the aforementioned problems that is broadly applicable.

2. Contributions

The main contribution of this work is the proposal of a mathematical formulation to deal with rhythmicity-related problems using ORI methodology. The key of this formulation is the definition of *circular signal*. Graphically, a *circular signal* can be mapped as a function displaying a temporal up-down-up pattern (see left panel in Figure 1), that underlies in many biological rhythmic processes (see left panels in Figure 2). These patterns, over a discrete number of values, can be defined using order restrictions as follows.

Definition 1 *Circular signal or up-down-up signal*

A signal $\boldsymbol{\mu}$ *in the Euclidean space is said to be circular iff* $\boldsymbol{\mu} \in C = \bigcup_{LU} C_{LU}$, *where* $L, U \in \{1, \ldots, n\}$, $C_{LU} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_1 \leq \cdots \leq \mu_U \geq \cdots \geq \mu_L \leq \cdots \leq \mu_n \leq \mu_1\}$ *if* $L > U$ *and* $C_{LU} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_1 \geq \cdots \geq \mu_L \leq \cdots \leq \mu_U \geq \cdots \geq \mu_n \geq \mu_1\}$ *if* $L < U$.



$$\phi_i = T_{LU}(\mu_i) = \begin{cases} \arcsin(\mu_i) - \frac{\pi}{2} & \text{, if } i \in \{1, \ldots, U\} \cup \{L, \ldots, n\} \\ \frac{\pi}{2} - \arcsin(\mu_i) & \text{, otherwise} \end{cases}$$

Figure 1: Equivalent formulation of circular signal.

To analyse observed data, we will consider a (circular) *signal* plus error statistical model. The appropriate statistical procedure to make inferences on those models formulated using restrictions, as the model proposed here, is Isotonic Regression (IR) [5]. IR is defined as the solution of a least squared minimization problem that looks for the best fit to a model incorporating restrictions among the parameters. Hence, IR provides an estimator for circular signal. To our knowledge, the IR problem for circular signals has not been studied in literature. Thus, another contribution of this work is the development and implementation, on the statistical software R, of a computationally efficient algorithm that computes IR for circular signals.

The IR circular signal estimator is the key to solve many rhytmicity-related problems such as rhythmiticy detection, which is formulated as an hypothesis testing problem contrasting a plain signal pattern against a circular signal. Our proposal to solve this test, that incorporates restrictions on the alternative hypothesis, is to conduct a conditional test [6], based on the maximum likelihood ratio test. Minor rhythmicity-related problems, such as peak time estimation or rhythm-pattern comparisons along different species, are likewise solved using the proposed ORI methodology.

Other interesting feature, thoroughly analysed in this work, that characterises circular signals is that they can be equivalently formulated within the Circular space (see Figure 1), from which its name is derived. The rigorous use of the circular geometry is fundamental to formulate and solve problems such as timing estimation when, due to experimental conditions, sample timing is unrecorded. The novel formulation allows us to define the model within the Circular space and to formulate the timing estimation problem as the problem of deriving the optimal circular order among the observed data. The latter problem can be solved as a minimization problem that is itself approximated by related Travelling Salesman Problem.

3. Results

Our new methodological proposals are validated and compared against popular alternatives in literature, which differ from one problem to the other, using simulated and real data. Due to space limitations, only two of those analyses are included. To assess the performance of ORI methodology detecting rhythmicity, an artificial database, that imitates what occurs in practice, is designed. It contains 15000 simulated patterns, of which 20% are rhythmic. The left group of patterns in Table 1 generates rhythmic genes and the right group does non rhythmic ones. From Table 1, ORI outperforms JTK on detecting, among others, genes that display asymmetric rhythmic patterns while controlling misclassification rates. Figure 2 displays the real (left panel) expression patterns of the genes *Per2* (top line) and *Per3* (bottom line) from NIH3T3 mouse liver cell lines [3] together with reordered patterns according to ORI (middle panel) and CYCLOPS (right panel) timing estimates, when the moment at which samples were taken is assumed to be unknown. From Figure 2, ORI estimates provide clearly rhythmic patterns closer to the real ones than CYCLOPS does.

Table 1: False Negative (FNR) and Discovery (FDR) Rates at nominal level of $\alpha = 0.01$.

FNR (for Rhythmic Patterns)						FDR (for Non Rhythmic Patterns)	
Cosine		*Cosine Two*		*Asymmetric*		*Flat*	
ORI	JTK	ORI	JTK	ORI	JTK	ORI	JTK
0.000	0.000	0.000	0.000	0.000	0.956	0.025	0.000

ORI methodology not only provides more efficient solutions to the rhythmicity-related problems described in this work, but also, promising expectations can be glimpsed on it. Due to its flexibility and versatility, we expect the ORI methodology to provide good solutions to many other chronobiological problems such as, for example, the analysis of rhythmic data in the Circular space including covariates in the model, which will be part of our future work.

Bibliography

[1] Larriba Y. *et al.* (2016). Order restricted inference for oscillatory systems for detecting rhythmic signals. *NAR* 44, e163.

[2] Larriba Y. *et al.* (2018). A bootstrap based measure robust to the choice of normalization methods for detecting rhythmic features in high dimensional data. *Front Genet* 9, 24.

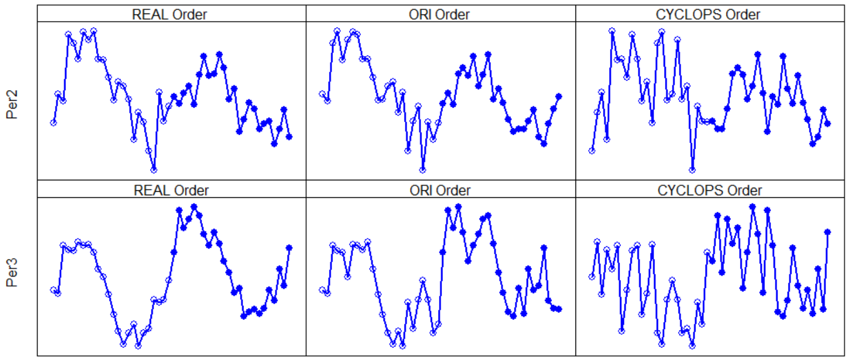


Figure 2: *Per2* and *Per3* gene-expressions plotted using three different orders.

[3] Hughes M.E., Hogenesch J.B., Kornacker K. (2010). JTK CYCLE: An Efficient Nonparametric Algorithm for Detecting Rhythmic Components in Genome-Scale Data Sets. *J Biol Rhythms* 25, 372–380.

[4] Anafi R.C. *et al.* (2017). CYCLOPS reveals human transcriptional rhythms in health and disease. *PNAS* 20, 5312–5317.

[5] Robertson T., Wright F.T., Dykstra R.L. (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, New York.

[6] Menéndez J.A., Salvador B. (1991). Anomalies of the likelihood ratio tests for testing restricted hypothesis. *Ann Stat* 19, 889–898.

The integrated nested Laplace approximation in order to fit Bayesian Dirichlet regression

Joaquín Martínez-Minaya[1], *Finn Lindgren*[2], *Antonio López-Quílez*[1], *Daniel Simpson*[3], *David Conesa*[1]

[1] Department of Statistics and OR, University of Valencia

[2] School of Mathematics, University of Edinburgh

[3] Department of Statistical Sciences, University of Toronto

Abstract

Dirichlet regression models can be used to analyze a set of variables lying in a bounded interval that sum up to one exhibiting skewness and heteroscedasticity, without having to transform the data. These data which mainly consist of proportions or percentages of disjoint categories are widely known as compositional data and are common in areas such as ecology, geology, and psychology. Bayesian inference has become a popular tool to deal with complex models such as Dirichlet regression models. The integrated nested Laplace approximation (INLA) is a widely extended methodology that for a large class of models provides higher accuracy for a limited computational budget. However, the implemented `R-INLA` package can not deal with multivariate likelihoods, such as, in particular, the Dirichlet likelihood. In this work, we propose an expansion of the INLA method for Dirichlet regression.

Keywords: Dirichlet regression, INLA, multivariate likelihood.

1. Introduction

Compositional data [1], consisting on proportions or percentages of disjoint categories adding to one, play an important role in many fields such as ecology, geology, etc. In the last years different approaches implemented in `R` packages have been presented in the Bayesian context: `BayesX`, `Bugs` or `R-JAGS` [2]. However, the integrated nested Laplace approximation (INLA) methodology does not allow to deal with compositional data when the number of categories is bigger than 2. In this work, we present a way to deal with this kind of data using INLA which has been implemented in the `R`-package `dirinla`.

The remaining of the document is structured as follows. Section 2 introduces basics of the Dirichlet regression. Section 3 gives a basic understanding about the INLA methodology. In section 4, the new approach is depicted followed by a simulation study performed to illustrate its good behavior along with a real example in section 5. Finally, section 6 concludes.

2. Hierarchical Dirichlet regression

The Dirichlet distribution is the generalization of the widely known beta distribution, and it is defined by the following probability density

$$p(\boldsymbol{y} \mid \boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{c=1}^{d} y_c^{\alpha_c - 1}\,, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ is known as the vector of shape parameters for each category, $\alpha_c > 0\ \forall c$, $y_c \in (0,1)$, $\sum_{c=1}^{d} y_c = 1$, and $\mathrm{B}(\boldsymbol{\alpha})$ is the multinomial beta function. The sum of all αs, i.e., $\alpha_0 = \sum_{c=1}^{d} \alpha_c$ is usually interpreted as a precision parameter. Beta distribution is the particular case when $d = 2$. To define the Dirichlet regression, let $\boldsymbol{Y}$ be a matrix with d rows and N columns denoting N observations for the different categories d of the response variable $\boldsymbol{y} \sim \mathcal{D}(\boldsymbol{\alpha})$. Then the model is set up as:

$$g(\alpha_{ci}) = \eta_{ci} = \boldsymbol{V}_{ci\bullet}\boldsymbol{\beta}_{\bullet c} \ , \tag{2}$$

where η_{ci} denotes the value of the linear predictor for the ith observation in the cth category, so $\boldsymbol{\eta}$ is a matrix with d rows and N columns. $\boldsymbol{V}$ represents a three dimensional matrix with dimension $d \times N \times J_c$ which contains the covariates values for each individual and each category, so $\boldsymbol{V}_{ci\bullet}$ shows the covariates values for the ith observation and the cth category. $\boldsymbol{\beta}$ is a matrix with J_c rows and d columns representing the regression coefficients in each dimension; and $g(\cdot)$ the link-function, in this case as $\alpha_c > 0\,, \forall c = 1, \ldots, d$, the $\log(\cdot)$ is employed.

3. INLA

The INLA methodology is now a well-established tool in Bayesian inference for latent Gaussian models (LGMs) [3] implemented in the R package R-INLA. The statistical inference is obtained using a three-stage hierarchical model formulation, in which observations $\boldsymbol{y}$ can be assumed to be conditionally independent, given a latent Gaussian random field (GF) $\boldsymbol{x}$ and hyperparameters $\boldsymbol{\theta_1}$, $\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta_1} \sim \prod_{i=1}^{N} p(y_i \mid x_i, \boldsymbol{\theta_1})$.

The versatility of the model class relates to the specification of the GF, $\boldsymbol{x} \mid \boldsymbol{\theta_2} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta_2}), \boldsymbol{Q}^{-1}(\boldsymbol{\theta_2}))$ which includes all the latent (nonobservable) components of interest such as fixed effects and random terms describing the underlying process of the data. The hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ control the GF and/or the likelihood for the data. When the precision matrix $\boldsymbol{Q}(\boldsymbol{\theta_2})$ is sparse, a GF becomes a Gaussian Markov random field (GMRF). When making inference with GMRFs, linear algebra operations are performed using numerical methods for sparse matrices, that jointly with the Laplace approximation, yields computational benefits.

Nevertheless, despite the advantages of R-INLA implementation, there are some limitations. For instance, R-INLA is not able to work with multivariate response variables, as it can not associate more than one data to only one individual. So, in order to fit the Dirichlet regression, we propose an extension of the paper [3] for models with multivariate response.

4. Inference. The new approach

As we have pointed out before, Dirichlet likelihood is intractable when models are fitted using R-INLA. In order to make this likehood handy, we propose convert it to independent Gaussian pseudo-observations which R-INLA can deal with. So, in this section, we approximate the log-likelihood function $\log p(\boldsymbol{Y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ using the Laplace approximation getting conditioned independent Gaussian pseudo-observations which are suitable to handle with R-INLA.

Let $\boldsymbol{\eta}_i := \boldsymbol{\eta}_{\bullet i}$ denotes the linear predictor corresponding to the ith observation $\boldsymbol{y}_i := \boldsymbol{Y}_{\bullet i}$, we define $l(\boldsymbol{y} \mid \boldsymbol{x}) = -\log p(\boldsymbol{y} \mid \boldsymbol{x})$ for any $\boldsymbol{y}$ and $\boldsymbol{x}$. In particular, we denote $l(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i) = -\log p(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i)$ the log-likelihood function expressed for the ith observation, being $\boldsymbol{y}_i$ and $\boldsymbol{\eta}_i$ vectors with d components. Using the Laplace approximation for $l(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i)$ in vector $\boldsymbol{\eta}_{\boldsymbol{0}_i}$, and defining $\boldsymbol{z}_{\boldsymbol{0}_i} := \boldsymbol{L}_{\boldsymbol{0}_i}^T(\boldsymbol{\eta}_{\boldsymbol{0}_i} - \boldsymbol{H}_{\boldsymbol{0}\boldsymbol{\eta}_i}^{-1}\boldsymbol{g}_{\boldsymbol{0}\boldsymbol{\eta}_i}) = \boldsymbol{L}_{\boldsymbol{0}_i}^T\boldsymbol{\eta}_{\boldsymbol{0}_i} - \boldsymbol{L}_{\boldsymbol{0}_i}^{-1}\boldsymbol{g}_{\boldsymbol{0}\boldsymbol{\eta}_i}$, a conditionally Gaussian approximation is constructed:

$$l(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i) \ \approx \ C + \frac{1}{2}[\boldsymbol{z}_{\boldsymbol{0}_i} - \boldsymbol{L}_{\boldsymbol{0}_i}^T\boldsymbol{\eta}_i]^T[\boldsymbol{z}_{\boldsymbol{0}_i} - \boldsymbol{L}_{\boldsymbol{0}_i}^T\boldsymbol{\eta}_i]\,, \tag{3}$$

i.e. $\boldsymbol{z}_{\boldsymbol{0}_i} \mid \boldsymbol{\eta}_i \sim \mathrm{N}(\boldsymbol{L}_{\boldsymbol{0}_i}^T \boldsymbol{\eta}_i, \boldsymbol{I_d})$ and $z_{0ik} \mid \boldsymbol{\eta}_i \sim \mathrm{N}([\boldsymbol{L}_{\boldsymbol{0}_i}^T \boldsymbol{\eta}_i]_k, 1)$, being $\boldsymbol{g}_{\boldsymbol{0}\boldsymbol{\eta}_i} = \nabla_{\boldsymbol{\eta}_i}(l)(\boldsymbol{\eta}_{\boldsymbol{0}_i}, \boldsymbol{y}_i)$ and $\boldsymbol{H}_{\boldsymbol{0}\boldsymbol{\eta}_i}$ the true hessian $(\nabla^2_{\boldsymbol{\eta}_i}(l)(\boldsymbol{\eta}_{\boldsymbol{0}_i}, \boldsymbol{y}_i))$. C is a constant whose value is $l(\boldsymbol{y}_i \mid \boldsymbol{\eta}_{\boldsymbol{0}_i}) - \frac{1}{2}\boldsymbol{g}_{\boldsymbol{0}\boldsymbol{\eta}_i}^T (\boldsymbol{H}_{\boldsymbol{0}\boldsymbol{\eta}_i}^{-1})^T \boldsymbol{g}_{\boldsymbol{0}\boldsymbol{\eta}_i}$ and $\boldsymbol{L}_{\boldsymbol{0}_i}$ is the result of applying the Cholesky factorization to $\boldsymbol{H}_{\boldsymbol{0}\boldsymbol{\eta}_i}$, $\boldsymbol{H}_{\boldsymbol{0}\boldsymbol{\eta}_i} = \boldsymbol{L}_{\boldsymbol{0}_i} \boldsymbol{L}_{\boldsymbol{0}_i}^T$. The observation vector $\boldsymbol{y}_i$ has been converted into Gaussian conditionally independent pseudo-observations $\boldsymbol{z}_{\boldsymbol{0}_i}$ that `R-INLA` can deal with. The approximation can be easily expanded to multiple observations. This methodology is being implemented in the package `dirinla`.

5. Simulation study and real example

In this section, we propose two examples in order to compare our approximation with a widely used method for Bayesian inference using MCMC algorithms, `R-JAGS` [2]. In both examples, we illustrate posterior distributions of the latent field $\boldsymbol{x}$. We assume that
$\boldsymbol{Y_i} \sim \text{Dirichlet}(\alpha_{i1}, \ldots, \alpha_{i4})\,, i = 1, \ldots, N$.

To the purpose of fitting the model with MCMC algorithms, the number of iterations used with `R-JAGS` has been 1000 with a burning of 100, thin 5 and 3 chains for the three different simulated datasets. On the contrary, in order to have a really good representation of the posterior, `long R-JAGS` has been performed using 1000000 of iterations with a burning of 100000, thin 5 and 3 chains.

5.1 Simulation

In this simulation, α_{ic} are just related with one parameter, i.e., $\log(\alpha_{ci}) = \beta_{0c}\ ,\ c = 1, \ldots, 4\,.$. A dataset with $N = 500$ using this structure has been simulated. The values β_{0c} with $c = 1, \ldots, 4$ have been -2.4, 1.2, -3.1, 1.3 respectively. In order to fit the model, some vague prior distributions for the latent field have been settled, in particular $p(x_i) \sim \mathrm{N}(0, \tau = 0.0001)$. In figure 1, the marginal posterior distributions for the β_{0c} with $1, \ldots, 4$ are displayed showing that the approximation performance is almost perfect comparing with models fitted using `R-JAGS` and with the real value.

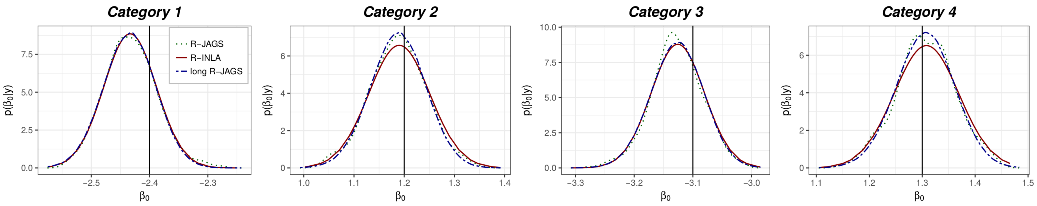


Figure 1: Simulation. Marginal posterior distributions of the latent field for the different categories, and using different methodologies `R-JAGS`, `R-INLA` and long `R-JAGS`.

5.2 Real example

The data of this real example has been extracted from [1]. The aim has been to do a pebble analysis of glacial tills. The total number of pebbles in each of 92 samples has been counted and the pebbles has been sorted into four categories: A (red sandstone), B (gray sandstone), C (crystalline) and D (miscellaneous). The percentages of these four categories and the total pebble counts have been recorded. The glaciologist has been interested in describing whether the compositions are in any way related to abundance. Here the linear predictor is composed by the intercept parameter and the slope corresponding to the total pebble counts. $\log(\alpha_{1i}) = \beta_{0c} + \beta_{1c} Pcount_i\ \ c = 1, \ldots, 4\,.$ Vague prior distributions for the latent field has been settled, in particular $p(x_i) \sim \mathrm{N}(0, \tau = 0.0001)$.

In figure 1, the marginal posterior distributions for the β_{0c} with $1, \dots, 4$ are displayed showing that the approximation performance is almost perfect comparing with models fitted using R−JAGS and with the real value.

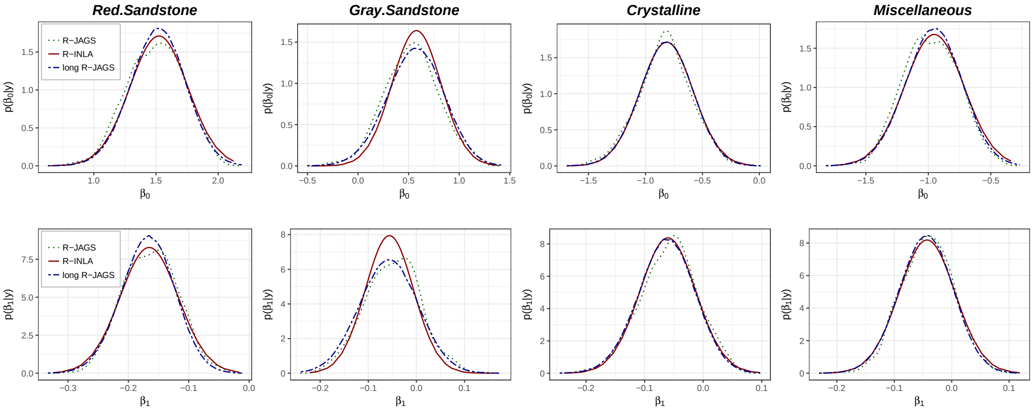


Figure 1: Real example. Marginal posterior distributions of the latent field for the different categories, and using different methodologies R−JAGS, R−INLA and long R−JAGS.

6 Conclusions

In this paper, the INLA methodology is extended to fit a model with a multivariate likelihood, the Dirichlet regression. The main idea is to approximate it by likelihoods that can be fitted by R−INLA, in this particular case, using a Gaussian likelihood. With regard to the computational aspect, here, we are presenting some results in order to fit models with just fixed effects. But there is still work to do. As we are converting original multivariate observations in conditionally independent Gaussian observations which only depends on the linear predictor, we expect to be able to incorporate random effects to the model, in particular, all the random effects which R−INLA can deal with allowing the user to fit spatial, temporal and spatio-temporal models.

Bibliography

[1] Aitchison, J., Egozcue, J. J., 2005. Compositional data analysis: where are we and where should we be heading? Mathematical Geology 37 (7), 829–850.

[2] Plummer, M., 2016. rjags: Bayesian Graphical Models using MCMC. R package version 4-6.

[3] Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (2), 319–392.

Contribuciones Orales

Smoothing methods for ‘very’ large data in spatio-temporal disease mapping: estimating mortality risks in Spanish municipalities

A. Adin[1,2], *T. Goicoa*[1,2], *M.D. Ugarte*[1,2]

aritz.adin@unavarra.es, tomas.goicoa@unavarra.es, lola@unavarra.es

[1] Department of Statistics, Computer Science and Mathematics, Public University of Navarre
[2] Institute for Advanced Materials (InaMat), Public University of Navarre

Many statistical models and computational methods have been developed during the last years to smooth mortality/incidence risks in disease mapping borrowing information from space and time. However, are these smoothing methods feasible when analyzing ‘very’ large spatio-temporal datasets?

In this work, we compare two different techniques to fit spatio-temporal disease mapping models based on both conditional autoregressive (CAR) structures, and multidimensional P-splines when the number of areas is very large. In particular we focus on the INLA (integrated nested Laplace approximation) approach and a scalable method for generalized additive models implemented in the `bam` function of the R package `mgcv`. The techniques will be used to analyze mortality caused by some of the more frequent cancer locations in 8000 municipalities in Spain during the period 1990-2015.

Keywords: Disease mapping, smoothing, massive data.

Cox elastic-net regularization for mRNA identification as predictors of HIV viral rebound

Yovaninna Alarcón-Soto[1], *Klaus Langohr*[1], *Guadalupe Gómez-Melis*[1]

[1]yovaninna.alarcon@upc.edu, Department of Statistics and Operational Research, Universitat Politécnica de Catalunya

A Cox Model regularized by an elastic net penalty has been considered to study the potential risk factors of HIV viral rebound. The motivation of the present work comes from a clinical trial on the efficacy of a specific therapeutic vaccine. Three different groups were compared: a) 3 vaccine doses and stop of the antiretroviral treatment (ART) at the first infusion ($n = 12$); b) 3 placebo doses ($n = 11$); and c) 3 vaccine doses and ART stop at the third infusion ($n = 12$).

The Cox Proportional Hazards model relates predictor variables $X = (X_1, \dots, X_p)^T$ and survival time T through the following model for the hazard function:

$$h(t|x) = h_0(t) \exp(x^T \beta), \tag{4}$$

where $h_0(t)$ is the baseline hazard and $\beta = (\beta_1, \dots, \beta_p)^T$. In this work, T is the time to viral rebound and the predictors X are messenger RNAs (mRNAs). As this work addresses the analysis of high-dimensional survival data, one of the problems is to select a submodel of (4) by providing a sparse estimate of β. Given a sample of n subjects, let T_i and C_i be the times to viral rebound and censoring, respectively, for subject $i = 1, \dots, n$. Write $Y_i = \min(T_i, C_i)$ and let the event indicator be $\delta_i = I(T_i \leq C_i)$. When a Cox elastic net is considered, to estimate β, we are solving

$$\min_{\beta} \sum_{i=1}^{n} \delta_i x_i^T \beta - \sum_{i=1}^{n} \delta_i \log \big(\sum_{j \in R_i} \exp(x_j^T \beta) + \frac{\gamma}{2} \sum_{j=1}^{p} \beta_j^2 + \lambda \sum_{j=1}^{p} |\beta_j| \big) \tag{5}$$

where $\gamma \geq 0$ and $\lambda \geq 0$ are two regularization parameters. Elastic net penalization, similarly to Lasso's regression, performs automatic variable selection and continuous shrinkage. In addition, such as ridge regression, elastic net deals with the problem of collinearity.

The use of the elastic net proportional hazards regression identified 4 mRNAs (SMC4, B3GAT1, BTG1, MYCN) as potential biomarkers for HIV viral rebound. These findings are clinically relevant since they can help to determine the effectiveness of the vaccine.

Keywords: Cox elastic net, transcriptomics, survival analysis.

Bootstrap-based methods for testing linear combinations of proportions. Application to length distribution samples of the cod in NAFO Division 3M.

Álvarez Hernández, M.[1], *Roca-Pardiñas, J.*[2] *and González-Troncoso, D.*[3]

[1]maria.alvarez@cud.uvigo.es, Centro Universitario de la Defensa - ENM
[2]roca@uvigo.es, Department of Statistics, University of Vigo
[3]diana.gonzalez@ieo.es, Instituto Español de Oceanografía - CO de Vigo

In recent years, inferences about a linear combination of proportions have aroused a great amount of interest, especially in applied research. By this way, in the present work studies were realized about the status of a fishing species in which, because of the type of sampling that is performed, the objective value is treated as a linear combination of independent proportions.

Many authors have developed procedures for making inferences (hypothesis test or confidence interval) about that parameter whose data are usually presented as a contingency table. As the exact methods are not viable for moderately large sample sizes, researchers have shown interest in asymptotic methods.

We introduce in this work a bootstrap method which is a competitive procedure and a good alternative vs methods proposed in literature. The bootstrap procedure additionally could allow to expand the findings in order to compare several linear combinations of independent proportions.

The methodology will be applied to real data of the cod in the Flemish Cap Bank (NAFO Division 3M). Biologists have assumed that the growth for adult males and females is the same in this stock, so length distributions are not sorted by sex. We must determine if the separation of the catch into sex is adequate and, in this case, if the differences between the proportion of sex are maintained over the years.

Keywords: Linear combination of proportions, bootstrap methods, cod.

A continuous-time hidden Markov model for the detection of hepatocellular carcinoma onset using serum biomarkers

Ruben Amorós[1,2], *Ruth King*[2], *Hidenori Toyoda*[3], *Takashi Kumada*[3], *Philip J Johnson*[4], *Thomas G Bird*[5,6]

[1]ruben.amoros@ed.ac.uk

[2] School of mathematics, University of Edinburgh

[3] Department of Gastroenterology, Ogaki Municipal Hospital

[4] Institute of Translational Medicine, University of Liverpool

[5] Cancer Research UK Beatson Institute

[6] Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh

Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer in adults which kills more than 700,000 globally per year, and early detection is essential for successful treatment. The use of serum biomarkers, such as alpha-fetoprotein (AFP) or the combination of several biomarkers, age and sex in the so-called GALAD score, have been proposed to detect the presence of tumours, with an increasing level indicative of potential cancer present. Previous static cut-off levels have been shown to be inefficient in detecting HCC due in part to the individual baseline heterogeneity, but an exploratory study suggests that analysing biomarker levels over time is a promising avenue for detecting the development of HCC. In this work we propose a Bayesian hierarchical model for longitudinal GALAD scores of patients under HCC screening to identify changes in the trend of the GALAD score, that can be indicative of the development of HCC. The hidden states correspond to the absence or presenc of HCC at the given time, with the later being an absorbent state. The model is additionally informed by the information on the diagnosis by standard clinical practice, taking into account that there may be false negatives within the diagnosis data. We apply the proposed model to a Japanese cohort database of patients under HCC surveillance and show that the detection capability of this proposal is greater that using a fixed cut-off point on the GALAD score.

Keywords: Longitudinal, change-point model, cancer detection.

Exploring the randomness of mentally generated head-tail sequences in healthy older adults

S.Baena-Mirabete[1], *S. Fernández Guinea*[2], *M.R. García-Viedma*[3], *P.Puig*[4]

[1]sergio.baena@uab.cat, Dept. of Mathematics, Universitat Autònoma de Barcelona
[2]sguinea@psi.ucm.es, Dept. of Experimental Psychology, Universidad Complutense de Madrid
[3]mrgarcia@ujaen.es, Dept. of Psychology, Universidad de Jaén
[4]ppuig@mat.uab.cat, Dept. of Mathematics, Universitat Autònoma de Barcelona

The analysis of categorical time series has been used to explore human memory patterns. Thus, for example, in an experiment of mentally tossing a coin, people believe, erroneously, that the coin alternates from heads to tails more often than really does. We present a study involving generation of *random* binary sequences by healthy older adults. We conducted an experiment in which individuals were asked to mentally simulate a fair coin. To that end, the subjects were each to produce a single sequence of 60 head-tail outcomes, simulating the behaviour of a fair coin, without seeing (but perhaps remembering) the past outcomes. The observational study presented in this talk is framed in the context of longitudinal data analysis in which a binary response, for a same individual, is repeated at 60 time points. A Markov chain of order (memory) k, taking values in a finite state space, is a well-known probabilistic model usually used to describe long memory processes (Baena-Mirabete and Puig, 2018). For more sophisticated processes, a combination of more than one Markov chains is a powerful alternative. In activities concerning the generation of random values by humans, one would expect to explain the deviation from randomness by a finite mixture that captures the population heterogeneity of the transition probabilities. In this talk we propose mixture models based on Markov chains from different approaches, as described in Baena-Mirabete and et al. (2019). The proposed models provide a tool that can be applied for early detection of Mild Cognitive Impairment and Alzheimer disease.

Keywords: Longitudinal data, higher-order Markov chains, mixture of Markov chains.

References

Baena-Mirabete, S. and Puig P. (2018) Parsimonious higher order Markov models for rating transitions, *J. R. Stat. Soc. A*, 181: 107-131. doi:10.1111/rssa.12267

Baena-Mirabete, S., Espinal, A., and Puig, P. (2019) Exploring the randomness of mentally head-tail sequences, *Statistical Modelling*, https://doi.org/10.1177/1471082X18816410

Co-ocurrence: modelling species simultaneously using a coregionalization approach

X. Barber[1], *D. Conesa*[2], *A. López-Quílez*[2], *M.G. Pennino*[3], *A. Esteban* [4]

[1]xbarber@umh.esm, Center for Operations Research, Universidad Miguel Hernández

[2] Departament d'Estadística i I.O., Universitat de València

[3] Instituto Español de Oceanogría, Centro oceanográfico de Vigo

[4] Instituto Español de Oceanogría, Centro oceanográfico de Murcia

One of the major goals of ecology is to understand the spatio-temporal dynamics of species. Commonly the habitat relevance is analyzed using Species Distribution Models (SDMs). These models link information on the presence/absence or abundance of a species to environmental variables to predict where (and how much of) a species is likely to be present in un-sampled locations or time periods.

However, biological interactions such as prey-predator, host-parasitic and/or adult-recruits can affect the species distribution in addition to environmental factors and need to be explicitly considered in process-based oriented modeling such as joint multi-species spatial models.

Here we present a Bayesian spatial joint species modelling to analyse the spatio-temporal dynamic of the adult-recruits relationship of the European hake (Merluccius merluccius) in the Iberian Mediterranean Spanish coastal areas. This statistical approach is a conditional corregionalized Bayesian linear model that makes use of a reparametrization of the variance to elicitate the prior distribution and the approximation of the posterior predictive distribution of the parameters in an easier way.

A posterior predictive distribution and a spatial probability distribution for each life-stage of the European hake were obtained.

We argue that this analytical framework allowed (1) to assess the spatial-temporal dynamic of the European hake adult-recruit relationship and (2) to provide an interesting tool in the context of multi-species modelling.

Keywords: Species distribution models, coregionalized models, multivariate Bayesian spatial models, fish habitat.

AMS: 62P12, 62M30

A bootstrap hyphotesis test to select the optimal categorization of the lymph node ratio for patients with colon cancer

Irantzu Barrio[1],*Javier Roca-Pardiñas*[2], *Inmaculada Arostegui*[3] *and CCR-CARESS Study Group*

[1]irantzu.barrio@ehu.eus, Departamento de Matemática Aplicada, Estadística e Investigación Operativa, UPV/EHU

[2]roca@uvigo.es, Departamento de Estadística e Investigación Operativa, Universidade de Vigo

[3]inmaculada.arostegui@ehu.eus, Departamento de Matemática Aplicada, Estadística e Investigación Operativa, UPV/EHU. BCAM- Basque Center for Applied Mathematics

Colorectal cancer is an important cause of mortality all over the world. To date, several risk prediction models for mortality and other adverse events have been developed to identify risk factors for this disease (Jorgensen ML, 2015). In this context, the prognostic impact of lymph node ratio (LNR), i.e. the ratio of metastatic to examined lymph nodes, is widely established (De Ridder et al., 2006; Rosenberg et al., 2008, 2010). In all the studies the LNR was categorized, however the cut points as well as the number of categories considered differed among the studies.

In the context of the logistic regression a methodology has been proposed to select the optimal cut points to categorize a continuous predictor based on the maximal AUC (Barrio et al., 2017). Although we can look for any possible number of cut points, in many circumstances the number of categories needed is unclear. In this work we propose a bootstrap hypothesis test to select the ideal number of categories. Specifically, given k the number of cut points, consideration will be given to a test for the null hypothesis for some k cut points $\mathbf{c}^0 = (c_1^0 < \ldots < c_k^0)$, versus the general hypothesis for some $k+1$ cut points $\mathbf{c}^1 = (c_1^1 < \ldots < c_{k+1}^1)$. To test H_0 we propose the use of a statistic based on the increment of the loss function $T = \widehat{AUC}(\hat{\mathbf{c}}^1) - \widehat{AUC}(\hat{\mathbf{c}}^0)$, where $\widehat{AUC}(\hat{\mathbf{c}}^0)$ and $\widehat{AUC}(\hat{\mathbf{c}}^1)$ are the estimated AUCs under the null and the general model respectively. We conducted a simulation study to evaluate the performance of T. The results suggest that the test performed well, with type I errors close to nominal errors.

Finally, we applied this methodology to a cohort study of patients with colon cancer in which we looked for the optimal cut-off points and number of categories of LNR. The results obtained were clinically validated.

Keywords: Categorisation, cut-off point, bootstrap.

Assessing surgical procedures using quality control charts

Blanco Alonso, Pilar[1], *Huerga Castro, M^{a}Carmen*[2]

[1]pilar.blanco@unileon.es, Dpto. de Economía y Estadística, Universidad de León

[2]carmen.huerga@unileon.es, Dpto. de Economía y Estadística, Universidad de León

The statistical control charts proposed by Shewhart have been traditionally used for controlling industrial processes. However, in various other contexts, there are processes that could be controlled by means of these methods. Such is the case of healthcare and hospital services, where procedures similar to a production process can be found.

In order to prove this rationale, in this paper we have selected one of the most frequent surgical procedures in public general hospitals, the so-called diagnosis-related group #162: *inguinal & femoral hernia procedures in adults (aged more than 17) without complications*. Since surgery duration is one of the most relevant variables in this procedure, it has been measured on the 445 operations performed in 2013 by ten different surgeons.

Firstly, a Kruskal-Wallis non-parametric proved the existence of significant differences in duration depending on the surgeon who performed the surgical intervention, so the global sample is split in 10 sub-samples, one per surgeon. Secondly, since control chart for individual measurements, which is the most appropriate one for controlling surgery duration, requires the data to be normal, a Lilliefors normality test was used to assess this assumption in each sub-sample, and control charts for individual measurements were applied to those sub-samples that met the normality assumption.

The resulting charts allow to monitor the stability of the surgery procedure, showing an estimate of the mean time of intervention and both upper and lower control limits at a three standard deviation distance from the mean. These limits can be considered as a benchmark for each surgeon's future interventions.

Keywords: Statistical quality control, control charts, healthcare, surgery procedures.

Seen once and more than once: a Bayesian approach to estimate the species richness

Anabel Blasco-Moreno[1], Pedro 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

Capture-recapture methods are used to estimate the species richness in Ecology, and they are also widely used in Social and Medical Sciences (Böhning et al., 2017).

Sometimes, the available data only consists in the total number of species observed (n) and the number of uniques (m), that is, the number of species observed exactly once. Therefore, $n - m$ is the number of super-duplicates (species observed two or more times). This is the case, for instance, of the study of the number of species in the coral reef of Tunku Abdul Rahman Marine Park by Chao et al. (2017). The main objective is to estimate the species richness, that is, the number of undetected species.

The likelihood function for this problem has a simple form,

$$L(X|p_0, p_1) = \left(\frac{p_1}{1-p_0}\right)^m \left(1 - \frac{p_1}{1-p_0}\right)^{n-m},$$

where p_0 is the proportion of undetected species (the parameter of interest), and p_1 is the proportion of species observed exactly once. Note that parameters are not identifiable and, consequently, a direct frequentist approach is not possible. However, Bayesian methods are suitable to cope with this problem.

Because $0 \leq p_0 + p_1 \leq 1$, we need to use prior distributions defined on the simplex $\mathcal{S}^3$. Dirichlet distributions (including the uniform) are not suitable because they act as "killer priors" leading to marginal posterior distributions for p_0 not depending on the observed data.

In this talk, several examples of application are analysed and discussed.

References

Böhning, D., van der Heijden, P. G. M. & Bunge, J. (2017). *Capture-Recapture Methods for the Social and Medical Sciences*. Chapman and Hall/CRC.

Chao, A., Colwell, R. K., Chiu, C. H., & Townsend, D. (2017). Seen once or more than once: Applying Good-Turing theory to estimate species richness using only unique observations and a species list. *Methods in Ecology and Evolution*, 8(10), 1221-1232.

Additive multidimensional disease mapping models: a case study on the Valencia city mortality

Paloma Botella-Rocamora[12], *Jordi Pérez-Panadés*[12], *Francisca Corpas-Burgos*[2], *Miguel A. Martínez-Beneito*[2]

[1]botella_pal@gva.es, Dirección General de Salud Pública - Conselleria de Sanitat Universal i Salut Pública
[2]Fundación para el Fomento de la Investigación Sanitaria y Biomédica (FISABIO) - Conselleria de Sanitat Universal i Salut Pública

Multivariate disease mapping models have become an important research field for biostatisticians and spatial epidemiologists for the last few years. These models deal with the joint geographical mapping of several diseases, taking into account their relationship. This approach allows obtaining, in addition to improved risk estimates, other interesting results as the correlation matrix between diseases.

In the same way that multivariate disease mapping models improve risk estimates, taking into account the relationships between other variables, such as sex, age group, time-period and so on, could also improve the risk estimates. These models, which include more than two variables (geographical units and at least two more variables) are called multidimensional disease mapping models.

A Bayesian multidimensional disease mapping framework has been proposed in the literature building their models trough a multiplicative structure for each of the variables considered. This proposal allows incorporating both structured (as time-period) and unstructured factors (such as disease or sex) in an easy manner. However, this proposal has some estimation problems when the number of factors included in the model increases and different variability parameters are included for different diseases. In this talk we propose an additive multidimensional reformulation that solves those estimation problems.

In this work we illustrate the performance of this proposal in the study of real mortality data for Valencia city. Specifically, we consider three factors in this study: geography (at a census tract level), time (5 time intervals between 1996 and 2015) and disease in the context of the MEDEA Project.

Keywords: Disease mapping, multivariate models, Bayesian analysis.

Detection of anomalies by gender in foot shapes

Ismael Cabero[1], *Irene Epifanio*[2], *Ana Piérola*[3]

[1]ismael.cabero@uv.es, Dept. de Didàctica de la Matemàtica, Universitat de València
[2]epifanio@uji.es, Dept. Matemàtiques-IMAC, Universitat Jaume I
[3]ana.pierola@ibv.org, Institut de Biomecànica de València

Knowledge of foot shape is of the great importance for the appropriate design of footwear. Not only a proper fit is a key factor in the decision to buy, but also poorly fitting footwear can cause foot pain and deformity, especially in women. Anthropometric studies devoted to apparel design do not usually carry out an outlier analysis, but it is fundamental, not only for data cleaning, but also to exploit the information that outliers can provide with regard to the design of shoes.

Therefore, we propose a new method for detecting outliers in multivariate data sets. It combines projections into relevant subspaces by archetype analysis with a nearest neighbor algorithm, through an appropriate ensemble of the results. A procedure for converting the outlier scores returned into binary labels is also proposed. The method has been assessed with 20 outlier detection algorithms with 8 benchmark data sets. The comparison shows that the performance of our proposal is very favorable. The new method is applied to detect outliers in foot shapes by gender.

Foot measurements are extracted from an anthropometric database of 775 3D right foot scans representing the Spanish adult female and male population. According to shoe design experts, we consider four variables: Foot Length (FL), Ball Girth (BG), Ball Width (BW), and Instep Height (IH). Outliers have been detected by the new method and interpreted for the sample of 382 women and 393 men. There are more outliers in women's feet than in men's. In the case of women, there are more outliers due to a very long FL than to a short FL; and as regards their shapes, many of the outliers are due to large dimensions in BG, BW and IH relative to their small FLs. However, for men there are fewer shape outliers but with more different typologies. In short, detecting the outliers can help shoe designers adjust their designs to a larger part of the population and be aware of the characteristics of the users that will make them uncomfortable to wear, whether when considering a range of special sizes or modifying any shoe feature to fit more customers.

Acknowledgments: This work is supported by grants DPI2017-87333-R from the Spanish Ministry of Science, Innovation and Universities and UJI-B2017-13 from Universitat Jaume I.

Keywords: Multivariate outlier detection, archetypal analysis, foot morphology.

A Bayesian longitudinal study of European sardine fishing

Gabriel Calvo[1], *Rubén Amorós*[2], *Carmen Armero*[1], *Maria Grazia Pennino*[3], *Luigi Spezia*[4]

[1]g.calvobayarri@gmail.com, Carmen.Armero@uv.es, Departamento de Estadística e Investigación Operativa, Universitat de València, Spain

[2]ruben.amoros.salvador@gmail.com, School of Mathematics, University of Edinburgh, UK

[3]graziapennino@yahoo.it, Instituto Español de Oceanografía, Vigo, Spain

[4]luigi@bioss.ac.uk, Biomathematics & Statistics Scotland, Aberdeen, UK

The European pilchard (Sardina pilchardus) is one of the most exploited small pelagic fish species which plays an important role in the transfer of energy from lower to higher trophic level organisms. However, some worrying changes have been observed in the European pilchard recently. Experts assure that the abundance of this fish has decreased. In fact, in 2017 the International Council for the Exploration of the Sea advised to stop fishing them for the next fifteen years everywhere.

In this work, we explore the temporal evolution of sardine fishing by means of Bayesian longitudinal modelling. The sardine dataset contains information by country about the quantity and the economic value of this kind of fish caught in the Mediterranean Sea from 1970 to 2014 as well as the fishing sector, gear type, area where the sardines were caught and catch type.

First, we approached this problem by means of linear mixed models with random effects over the intercept and slope. Then, we considered more complex models which include serial correlation such as autoregressive, moving average and non-linear models based on spline basis functions.

Finally, the presence of sudden changes in the trends of sardine fishing for some countries suggests the inclusion of latent variables in the form of mixture and hidden Markov models is worth future research.

Keywords: Fishery estimation, latent variables, serial correlation.

Age-space-time analysis of ovarian cancer mortality in Spain

Paula Camelia Trandafir[1,2], *Aritz Adin*[1,2] *and María Dolores Ugarte*[1,2]

camelia@unavarra.es, aritz.adin@unavarra.es, lola@unavarra.es

[1] Department of Statistics, Informatics and Mathematics, Public University of Navarre

[2] INAMAT, Public University of Navarre

Ovarian cancer is a leading cause of death from gynecological malignancy. Its highest incidence is reported in women aged 60-64 years, with most of the diagnoses occurring in women over 50. Every year there are 204 449 estimated new cases of ovarian cancer diagnosed worldwide (about 4% of cancers in females) with 124 860 disease-related deaths [1]. Within the epidemiological literature, one finds that many studies only consider spatial, temporal or spatio-temporal analysis of ovarian cancer mortality, without taking into account disaggregation by age groups. On occasions, this can lead to somewhat misleading conclusions. Our goal here is to study the temporal evolution of the geographical patterns of ovarian mortality rates by age groups in Spanish provinces during the period 1989-2015. Different autoregressive models will be considered for this task. Model fitting and inference will be carried out using integrated nested Laplace approximations [2] using the code developed by [3].

Keywords: CAR models, disease mapping,ovarian cancer.

Bibliography

[1] Bray F, Ferlay J, Soerjomataram I, Siegel R L, Torre, L A, Jemal, A. (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, **68**, pp 394-424.

[2] Rue H, Martino S. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations *Journal of the Royal Statistical Society*, **71(2)**, pp 319-392.

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

A Bayesian epidemiological compartmental model for labour (in)stability in Spain

Juan A. Carbonell[1], *Francisco Jose Santonja*[2], *Carles Xavier Simo*[3]

[1]carajuan@alumni.uv.es, Departament de Sociología i Antropología Social, Universitat de València.
[2]francisco.santonja@uv.es, Departament d'Estadística i Investigació Operativa, Universitat de València
[3]carles.simo@uv.es, Department of Sociology and Social Anthropology, Universitat de València.

Compartmental models are widely used in the study of infectious diseases dynamics. This type of models may be used to predict the prevalence (total number of infected) or for understanding what the most efficient strategy in order to define a vaccination plan. This approach suposses that population is divided into compartments depending on their status with respect to the disease and allows us to quantify the transitions between compartments.

In this work we present a (susceptible-infected-susceptible) SIS model to Spanish population born in 1960 and in 1970 in order to quantify transition between employed and un-employed population. Studies estimate these parameters with time invariant methods such as least squares, maximum likelihood and, recently, by means of hierarchical Bayesian modelling. Here we show a methodological approach in which paramater estimation is time dependant.

Results show that the speed from employment to unemployment increases drastically in the first years until it reaches a top asymptote and after eight years it slowly decreases to a bottom asymptote. On the other hand, speed from unmployment to employment shows an increase until a top asymptote in which it keeps constant. We anticipate our work to be a starting point to more sophisticated research in dynamic compartment models with the inclusion, for instance, of other functions or mixtures to improve fit or estimation techniques that take into account information from previous blocks.

Keywords: Compartment models, bayesian statistics, labour trajectories.

Technical performance variability of imaging biomarkers during acquisition of image

Jose Miguel Carot[1], *Andrea Conchado*[2], *Claudio Castro*[3], *Ángel Alberich-Bayarri*[4], *Luis Martí-Bonmatí*[5]

[1]jcarot@eio.upv.es, Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València

[2]anconpei@eio.upv.es, Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València

[3]ccastro@uv.mx, Universidad Veracruzana (México)

[4]alberich_ang@gva.es, Biomedical Imaging Research Group, Instituto de Investigación Sanitaria La Fe, QUIBIM SL, Valencia, Spain

[5]luis.marti@uv.es, Department of Radiology, Hospital Universitari i Politècnic La Fe

Recent technological advances in digital medical image and computational power of computer systems have enhanced the ability to diagnose and monitor diseases. This improvement relates to the extraction of quantitative information from images, which usually cannot be depicted through visual inspection. Imaging parameters will allow researchers for creating and developing new biomarkers if proven surrogate to treatment allocation or clinical endpoints. The creation and use of a new biomarker involve a previous in-depth study of the acquired images, including technical validation and uncertainties analysis. The acquisition stage validation is the first limiting step in this developing process, since data quality is a key factor. These results are essential to properly evaluate how acquired images, depending on different factors, may change expected results.

This research analyses and compares the effect of different acquisition factors, stored in the DICOM (Digital Imaging and Communication in Medicine) headers, on quality measurements such as signal-to-noise ratio (SNR) and background homogeneity. With this aim, we applied different statistical methods for the assessment of these technical performance, such as Measurement Systems Analysis (MSA), multivariate analysis, and visualization techniques.

Keywords: Imaging biomarkers, multivariate analysis, measurement systems analysis.

Comparison of spatio-temporal models with Matlab and R. Study of European for ages after retirement.

Patricia Carracedo[1], *Ana Debón*[2]

[1]patricia.carracedo@campusviu.es, Área de ciencias sociales, Universidad Internacional de Valencia
[2]andeau@eio.upv.es , Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València

During the 20th century, insurers and Social Security administrations have focused their attention on increasing life expectancy in Europe. These agents have information about the data collected in dynamic life tables for each countries. With the aim of complementing this information, this study calculates panel data models using MATLAB and R software. These models take into account the temporal and spatial dependence of the data and identify the covariables which explain mortality. A comparison between both softwares is established, detailing the steps to select the best model obtained. The selected models were validated by the measures of goodness of fit and tests over residuals. Mortality was quantified with the Comparative Mortality Figure which is the most suitable statistic for comparison of mortality by sex over space when detailed specific mortality is avalaible for each studied population. The panel data used corresponds to the male and female mortality of aged population between 65 and 110+ years of 26 European countries for the period 1995-2012. The covariates considered in the model were gross domestic product growth rate, health expenditure, CO2 emissions and education expense.

Keywords: Mortality, Europe, spatial panel data models.

Multivariate contributions to the mining of textual data, program in R MTMM, Case study, social problems in Mexico

Claudio Castro[1], *José Miguel Carot*[2], *Luis Duran*[3]

[1]ccastro@uv.mx, Centro de Estudios de Opinión y Análisis, Universidad Veracruzana
[2]jcarot@eio.upv.es, Centro de Gestión de la Calidad y el Cambio, Universidad Politécnica de Valencia
[3]jesus@gmail.com, Centro de Estudios de Opinión y Análisis, Universidad Veracruzana

Text mining is multidisciplinary, takes in consideration areas such as data mining, linguistics, computational statistics and information technology, using statistical methods such as text classification techniques, conglomerates, ontologies (creation of taxonomies) and latent corpus analysis. This paper presents a methodological proposal focused on the creation of a text mining module focused on multivariate textual analysis, implementing five statistical techniques (Cluster Analysis, Factorial Analysis, Simple and Multiple Correspondence Analysis, Factorial Analysis by Correspondence), with which one can find discursive patterns of opinion regarding a corpus (set of texts). In particular, this work focuses on describing in a general way each technique proposed in the implementation of the module in R of the program that we call Text Mining Module and Textual Ultitive Analysis (MTMM), which allows us to perform multivariate textual analysis of a large volume of information, this modular interface is compatible with 3 operating systems (Windows, MacOS and Linux) to perform multivariate analysis of texts. The analyzes carried out are: Law of Zipf (Montemurro, 2001), word clouds, analysis of similarity (Marchand & Ratinaud, 2012), factorial analysis of correspondence (Benzécri 1973) and classification analysis Reinert (Vizeu Camargo & Justo, 2013), which have different uses in areas such as qualitative research work, social representations, survey analysis, among others.

Likewise, an application is presented to the context of the study of citizen opinions of social problems in an area of the Mexican state of Veracruz.

Keywords: Multivariate analysis, text mining, opinion study.

Spatial Bayesian modelling applied to the surveys of *Xylella fastidiosa* in Alicante

M. Cendoya[1], *J. Martínez-Minaya*[2], *V. Dalmau*[3], *A. Ferrer*[3], *D. Conesa*[2], *A. López-Quílez*[2], *A. Vicent*[1]

[1]m.cendoya.m@gmail.com, Centre de Protecció Vegetal i Biotecnologia, Institut Valencià d'Investigacions Agràries (IVIA).

[2]Departament d'Estadística i Investigació Operativa, Universitat de València, Burjassot 46100.

[3]Servei de Sanitat Vegetal, Conselleria d'Agricultura, Medi Ambient, Canvi Climàtic i Desenvolupament Rural.

The increasing development of techniques applied to species distribution models (SDMs) has allowed them to be widely used in different areas such as ecology and epidemiology. The SDMs are useful tools to establish suitable conditions for the expansion of populations, to evaluate the associations of biotic and abiotic factors with the geographic extent of the species, as well as to predict the species distribution in space and in time. These types of models can be developed through different methodologies. However, many of them ignore the spatial dependence which usually exists among the geographical locations of the observations. This can lead to an overestimation of the parameters and establish false relationships between observations and covariates. Spatial Bayesian hierarchical models allow the inclusion of spatial autocorrelation. Here, this type of modeling was used to analyse the effect of climatic and spatial factors in the distribution of the bacterium *Xylella fastidiosa* subsp. *multiplex* in Alicante (Spain). The presence/absence data of *X. fastidiosa* were obtained from the official surveys gathered in 2017 in the demarcated area of Alicante, where the pathogen was detected affecting almond crops. Climatic covariates were obtained from the WorldClim database. The spatial effect was incorporated through a conditional autoregressive structure (iCAR), while a categorical variable was included based on the "Purcell's" levels of disease severity based on minimum winter temperature. These thresholds were defined in North America for Pierce's disease of grapevine, caused by *X. fastidiosa* subsp. *fastidiosa*, and it is unknown if they can be extrapolated to other subspecies or geographic regions. The Integrated Nested Laplace Approximation (INLA) method was used to obtain the posterior distributions of the model parameters. The results show that climatic covariates were not relevant in the model, probably due to the reduced geographic extent of the study area and therefore low climatic variability. However, the pathogen was detected in all "Purcell's" winter temperature thresholds, confirming the climatic adaptability of *X. fastidiosa* subsp. *multiplex*. In addition, the strong spatial effect observed indicated that the spatial structure has a central role on the dynamics of disease spread.

Keywords: Hierarchical Bayesian models, INLA, *Xylella fastidiosa*.

Latent growth curve models for analysing body image over time among women with breast cancer

Conchado, A.[1], *Marco, J.H.*[2], *Castejón, J.*[3]

[1]anconpei@eio.upv.es, Department of Applied Statistics and Operational Research, and Quality, Universitat Politécnica de València

[2]Jose.H.Marco@uv.es, Psychological, Personality, Evaluation and Treatment Department, Universitat de València

[3]Clínica de Psicología, Universidad Internacional de València

The application of Latent Growth Mixture Curve models (LGMC) to a longitudinal dataset concerning body image among women with breast cancer allowed us to distinguish between different groups of patients, according to their trajectories during a one - year period. Two different scales were used for measuring and testing body image: The Body Image Scale (BIS), and the Body Appreciation Scale (BAS). We initially collected responses from 115 patients in the moment before surgery, but some cases were lost throughout the period of study: post - surgery, 3 months, 6 months and 9 months - one year, thereby leaving 80 complete - data cases. We addressed subject attrition during this one - year period adding the corresponding specification of the model using the EQS model.

Our findings showed that women suffering breast cancer could be classified into three clusters regarding measures of body image via BIS scale. Two out of this three classes reached their maximum level three months after the beginning of treatment, consistently with conventional medical treatments including chemotherapy. Among them, the majority of women (98.2%) presented a standard body image of themselves over time, whereas two minor groups (0.9%, each of them) followed different trajectories. Multinomial regression was performed to the resulting latent classes using a series of preselected baseline variables (Age, Copying Style, Hopelessness, Depression, and Meaning in life) in order to determine if there might be differentiation of classes. As a result, any variable was found to predict significantly theses latent classes. However, the Kruskal-Wallis test indicated that the Coping Style, specifically the Anxious Worry during the conventional medical treatment, was higher in the two groups with higher scores in the BIS when compared to the group with moderate BIS scores.

Consistently with these results, two groups were proposed through the application of LCGM models to the BAS scale. The majority of patients (91.9%) presented a stable trajectory over time, with a slight decrease of BAS values three months after surgery. Logistic regression was performed to the resulting latent classes using a series of preselected baseline variables (Age, copying strategies, Hopelessness, Depression, and Meaning in life) in order to determine if there might be differentiation of classes. Only Depression significantly predicted the latent class of BAS. Both classifications of patients, corroborate our previous knowledge about psychopathological symptoms and other psychological information for patients classified in minor groups. These findings have important implications as regards to decisions concerning medical treatment planning.

Keywords: Breast cancer, body image, latent growth curve models, trajectories, time.

An adaptive CAR proposal to model spatial dependence in disease mapping studies

Francisca Corpas-Burgos[1], *Miguel A. Martínez-Beneito*[2]

[1]corpas_fra@gva.es, [2]martinez_mig@gva.es

[1,2]Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana (FISABIO)

Disease mapping pursues the study of the geographical distribution of health-related events, such as mortality, in order to identify those locations that show a higher risk. In the analysis of data aggregated by small geographical areas it is important to take into account the spatial dependence that exists between the areas in order to obtain more reliable estimates of the risks.

In the Bayesian literature, disease mapping models are frequently specified as generalized linear models that incorporate spatial dependence between nearby areas in the linear predictor through random effects following some spatial prior distribution. Surely, the most popular spatial prior distributions are the Conditional Autoregressive (CAR) distributions. CAR distributions induce spatial dependence by defining a neighborhood structure accounting for the geographical arrangement of the spatial units. That neighborhood structure is summarized by a spatial weights matrix quantifying the influence that the random effects of the areas have on each other, so those weights should reflect the strength of dependence between any two spatial areas.

By far, the most common procedure for defining the weight matrix in CAR distributions is to use an adjacency criterion. In that case all pairs of areas with adjacent borders are given the same weight, typically 1, and the rest of non-adjacent areas are assigned a weight of 0, reflecting independence given the rest of areas. However, this criterion assumes equal weights for all adjacent areas, what could be somewhat unrealistic and arbitrary in some settings.

The objective of this work is proposing an alternative specification of neighborhood matrix for CAR distributions in which spatial weights are considered as random variables in the model and estimated according to the data information. We will ilustrate the use of our adaptative proposal in a multivariate context for defining weights matrices which reflect the geographical performance of diseases in the region of study. The weight structure estimated would allow to derive improved risk estimates incorporating the empirical spatial dependence structure observed in that region.

Keywords: Spatial dependence, adaptive CAR models, disease mapping.

Antedependence skew-normal linear models for longitudinal data

Martha Lucía Corrales-Bossio[1], *Edilberto Cepeda-Cuervo*[2]

[1]martha.corrales@usa.edu.co, Department of Mathematics, Sergio Arboleda University

[2]ecepedac@unal.edu.co, Department of Statistics, National University of Colombia

In longitudinal data analysis, the assumption of multivariate normality may be questionable, especialy 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 descompositition of variance covariance matrix (Macchiavelli & Moser 1997, Cepeda 2001, Cepeda & Gamerman 2005), we propose joint modeling of the localization, scale, autoregressive and skewness, assuming skew-normal distributions of Azzalini and Sahu, for analysis of growth and development dataset (Corrales-Bossio & Cepeda-Cuervo 2017). The data of growth and development of deaf students in Colombian were soported by the National Institute for the Deaf (INSOR) in Bogota - Colombia. The studies of growth and development of the children are very important in clinical research, because these allow to detect of problems in their development.

Keywords: Longitudinal data, skew-normal.

Bibliography

Azzalini, A. & Capitanio, A. (1999), 'Statistical applications of the multivariate skew normal distribution', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(3), 579-602.

Azzalini, A. & Dalla Valle, A. (1996), 'The multivariate skew normal distribution.', *Biometrika* **83**(4), 715-726.

Cepeda, E. (2001), 'Modelagem da variabilidade em modelos lineares generalizados', *Unpublished Ph. D. tesis. Instituto de Matematicas. Universidade Federal do Rio do Janeiro.////http://www. docentes. unal. edu. co/ecepedac/docs/M http://www. bdigital. unal. edu. co/9394* **2**.

Cepeda, E. & Gamerman, D. (2005), 'Bayesian methodology for modeling parameters in the two parameter exponential family', *Revista Estadistica* **57**(168-169), 93-105.

Corrales-Bossio, M. & Cepeda-Cuervo, E. (2017), 'Skew-normal regression models. bayesian joint modeling of location, scale and shape parameters', *Working paper.*

Lin, T.-I. & Wang, Y.-J. (2009), 'A robust approach to joint modeling of mean and scale covariance for longitudinal data', *Journal of Statistical Planning and Inference* **139**(9), 3013-3026.

Macchiavelli, R. E. & Moser, E. B. (1997), 'Analysis of repeated measurements with antedependence covariance models', *Biometrical journal* **39**(3), 339-350.

Sahu, S. K., Dey, D. K. & Branco, M. D. (2003), 'A new class of multivariate skew distributions with applications to bayesian regression models', *Canadian Journal of Statistics* **31**(2), 129-150.

Analysing stability of the human microbiome by a Dirichlet autoregressive model

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

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

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

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

Recent studies suggest that microbiota, which is the collection of all bacteria living either in or on the human body, plays a key role in defining the health status of individuals. Several works have pointed out that the maintenance of a stable microbial ecosystem is necessary for a healthy life. In fact, it is known that a disruption of the stable state of the microbiota can be related to obesity, diabetes or cancer. Therefore, analyzing stability of microbiota and understanding how quickly it recovers and reaches a new stable state are key questions.

Many proposals for microbiota data longitudinal analysis are based on counts strategies. However, we suggest here to consider compositional vectors of relative abundances. Whitin this paradigm, longitudinal microbiota analyses are carried out taking into account multivariate time series which, at each time point, can be considered as a vector of non-negative proportions that sum to one. In this scenario, and assuming that relative abundances are distributed with a Dirichlet distribution with time-varying parameters, we propose an autoregressive model which enables analyzing the stability of microbiota time-series data.

Keywords: Microbiota, Dirichlet distribution, multivariate time-series.

A functional diagnostic test: how to classify an individual as healthy or disease from a functional biomarker?

Graciela Estévez-Pérez[1], *Philippe Vieu*[2]

[1]graci@udc.es, Departamento de Matemáticas, Universidade da Coruña, Spain

[2]philippe.vieu@math.univ-toulouse.fr, Institut de Mathématiques, Université Paul Sabatier, Toulouse, France

Nowadays, there are many fields, including Economy, Biology and Medicine, that handle functional data, that is, realizations of random elements taking values in an infinite dimensional space. As the analysis of this type of data requires using specific tools to extracting relevant information from them, numerous statistical techniques have been extended to functional context in the two last decades. However, the development of diagnostic methods and analysis of ROC curves for functional biomarkers is an exception. In fact, the FDA (functional data analysis) community has started debating recently on this topic but, at our knowledge, only when functional data are involved as covariables. The reason for this lack of development for ROC analysis with functional sample is probably linked with the difficulty of establishing some ordering in infinite-dimensional spaces. However, the technological advances are allowing to observe markers with more complex structures and the need for such a functional extension of ROC methodology becomes each day more important in the clinical practice and the medical research.

The main purpose of this work is constructing a diagnostic test based on a functional biomarker. It has been possible because of the new general methodology, developed by the authors, for ordering data in infinite dimensional spaces. Specifically, we construct a classification rule, we analyze precisely the concepts of specificity and sensitivity and we discuss how a functional version of ROC curve can help in balancing the trade-off between sensitivity and specificity. Finally, we propose a fully automatic diagnostic testing procedure in which the parameters of the method must be data-driven selected.

In addition, we present some results of a simulation study in R conducted to evaluate the performance of the proposed diagnostic test and to solve the computational issues linked with the choices of the various parameters of the methodology. We also implement some criteria to select the optimal cutoff curve and analyze the accuracy of functional diagnostic test proposed. Finally, the proposed procedure is shown on a real data set that comprise the expression of 100 genes in 25 tumor prostate samples and 25 normal samples.

Keywords: Test diagnostic, functional data, ROC curves.

Mixture-based clustering for ordinal data. A Bayesian approach

Daniel Fernández[1], *Richard Arnold* [2], *Shirley Pledger* [3]

[1]df.martinez@pssjd.org, Institut de Recerca Sant Joan de Déu, Parc Sanitari Sant Joan de Déu, CIBERSAM, Barcelona, Spain

[2]richard.arnold@vuw.ac.nz, School of Mathematics and Statistics, Victoria University of Wellington, New Zealand.

[3]shirley.pledger@vuw.ac.nz, School of Mathematics and Statistics, Victoria University of Wellington, New Zealand.

Many of the methods that deal with clustering in matrices of data are based on mathematical techniques such as distance-based algorithms or matrix decomposition. In general, it is not possible to use statistical inferences or select the appropriateness of a model via information criteria with these techniques because there is no underlying probability model. Additionally, the use of ordinal data is very common (e.g. Likert or pain scale). Recent research has developed a set of likelihood-based finite mixture models for a data matrix of ordinal data. This approach applies fuzzy clustering via finite mixtures to the stereotype model. Fuzzy allocation of rows, columns, and rows and columns simultaneously (biclustering) to corresponding clusters is obtained by performing a Reversible-Jump MCMC sampler. Examples with ordinal data sets will be shown to illustrate the application of this approach.

Keywords: Clustering, ordinal responses, RJMCMC.

Misreported diagnosis of the Attention Deficit Hyperactivity Disorder

Amanda Fernández-Fontelo[1], *Alejandra Cabaña*[2], *David Moriña*[2], *Anna Giménez Palomo*[3], *Pedro Puig*[2]

[1]fernanda@hu-berlin.de, Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin

[2]acabana@mat.uab.cat, david.morina@uab.cat, ppuig@mat.uab.cat, Departament de Matemàtiques, Universitat Autònoma de Barcelona

[3]agimenezp@clinic.cat, Department of Psychiatry. Institut Clínic de Neurociències. Hospital Clínic de Barcelona

Attention deficit hyperactivity disorder (ADHD) is a mental disorder usually diagnosed for the first time in school-aged children. Although ADHD is systematically associated with children and pre-adolescents, adults are also susceptible to have such behavioral disorder. On top of that, many studies support that ADHD symptoms vary from gender, culture, doctor criteria, socioeconomic level, among others.

Many researchers in the area point out that this mental disorder is strongly misdiagnosed since the criteria for diagnosis are still diffuse, and symptoms might dramatically differ depending on several characteristics (age, gender, etc.).

The present work is aimed at quantifying the degree of misreporting (or sometimes misdiagnosis) in different settings of the official TDAH records on a region of the province of Barcelona (Spain). In order to carry it out, the models proposed by Fernández-Fontelo et al.(Under-reported data analysis with INAR-hidden Markov chains. Statistics in Medicine; 2016, 20: 4875-4890) and FernÃ¡ndez-Fontelo et al. (Untangling serially dependent under-reported count data from gender-based violence; under review) are extended by introducing a new operator, called fatteningthinning operator, which is able to accommodate in such models both the phenomena of underreporting (or under-diagnosis) and over-reporting (or over-diagnosis).

Considering the new model based on the fattening-thinning operator, the following hypothesis are thoroughly studied: (H1) Do the misreporting issue differ from males and females?, (H2) Do the misreporting phenomenon vary between children, adolescents and adults?, and (H3) Is the interaction between age (children, adolescents and adults) and gender crucial to identify different misreporting patterns?

Results regarding hypothesis H1-H3 are presented and discussed, and significant conclusions are derived.

Keywords: Misreporting, ADHD, count time series.

Estimating health conditions of Andalusian healthcare professionals using online surveys with Propensity Score Adjustment

Ramón Ferri-García[1], *María del Mar Rueda*[2], *Andrés Cabrera-León*[3]

[1]rferri@ugr.es, Department of Statistics and Operations Research, University of Granada
[2]mrueda@ugr.es, Department of Statistics and Operations Research, University of Granada
[3]andres.cabrera.easp@juntadeandalucia.es, Andalusian School of Public Health (EASP), CIBER of Epidemiology and Public Health (CIBERESP), Instituto de Investigación Biosanitaria (ibs.granada)

Healthcare professionals are often subject to high work pressure and physical efforts, which eventually influence their levels of self-perceived health and life satisfaction. These indicators and their implications on public health have gained research attention over the years. The measurement of such aspects is usually studied using surveys which are not taken under a probabilistic scheme: they are filled by self-selected volunteers from the healthcare professionals' population, conforming a non-probabilistic sample whose selection and representation bias could lead to unreliable estimates, given that health and life satisfaction are not likely to be homogenous.

One of the approaches for inference in non-probabilistic samples is Propensity Score Adjustment (PSA). This adjustment is based on reweighting volunteer samples considering the probability of each individual to participate in the survey, by adjusting a propensity model on the combination of the volunteer sample and a probabilistic reference sample taken from the same target population.

In this work, PSA is applied in the analysis of a volunteer sample of 1,797 healthcare professionals from the region of Andalusia (Spain). PSA weights are computed with both logistic regression and Machine Learning algorithms, using data from the complete census of Andalusian healthcare professionals as the reference sample. These weights are applied on the estimation of life satisfaction and health-related variables, including self-perceived health. Results show that adjusted estimates can differ substantially from non-adjusted ones, suggesting that the presence of some patterns in the healthcare professionals population could be different from those provided by usual estimations.

Keywords: Propensity Score Adjustment, machine learning, healthcare professionals.

AMS: 62D05

Detecting violations to the conditional independence assumption in joint modelling of longitudinal data and time to the event

Alberto García-Hernandez[1], *Teresa Pérez*[2], *María del Carmen Pardo*[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

Content: Joint models (JM) of longitudinal and survival data have been extensively studied both in the survival field and in the missing data framework. In particular, our focus was the use of JM to analyse a longitudinal response adjusting for informative dropouts, that is, our time-to-event variable is the missingness of the longitudinal response. Joint models are grounded on the conditional independence (CI) assumption, that is, given the random effects, the longitudinal response and the event process are assumed independent. We have studied a violation to this assumption due to association between the event and the longitudinal response involving both the (latent) subject effect and the observational error.

Objectives: a) To quantify the effect of this deviation to the longitudinal response estimates; b) To compare JM with mixed models, using model selection techniques.

Methods: We have simulated data of 500 clinical trials under two possible mechanisms: a) assuming CI (no violations of the JM); b) assuming violation of the CI due to the association between the time-to-event and the longitudinal response involving the observational error. We have first considered a set of JM fitted using a likelihood method approach [1] and have obtained the fit of the survival component of the model following the method of Zhang et al. in which the global log-likelihood function is decomposed providing separate assessments of each component of the JM (LL$_{Joint}$ = LL$_{Long}$ + LL$_{Surv|Long}$) [2]. We have compared the fit provided by JM on the survival process (LL$_{Surv|Long}$) versus a set of comparable parametric survival models using the longitudinal response as a time-dependent-covariate.

Results: JM provided biased estimates when the missingness process depended evenly on both the (unobserved) subject mean and the observational error. Model selection methods distinguished well the studied violation from the situation where the conditional independence assumption held, and no other violation was presented.

Keywords: Joint modelling, model selection, log-likelihood function.

Surveys on the use of time with a gender approach, considering paid and unpaid work

Giorgini Diana[1], *Escanes Viviana*[2], *Filippini O. Susana*[3]

[1]dgiorgin@agro.uba.ar, Statistical Department, National University of Luján; Faculty of Agronomy, UBA
[2]vescanes@gmail.com, Social Science Department, University of Luján
[3]sfilippini@unlu.edu.ar, Statistical Department, National University of Luján; Faculty of Agronomy, UBA

The economic and social functioning of our societies and their eventual development depends directly on the activities that individuals perform in our daily lives. Both men and women differently distribute our time according to various activities such as paid work, unpaid work, recreational and personal care activities, leisure time, but with different proportions of time according to the type of membership. Historically, the work developed by women within the home has been socially and economically devalued. However, the evolution of gender studies in recent decades has made it possible to raise such problems in the public sphere. With the aim of generating better statistics on paid and unpaid work in our academic community, and producing information under certain standards, which allows comparison with other university areas, the project "Paid and unpaid work: Towards measurement was carried out. of the use of time in UNLu teachers". We proceeded to survey the work of teachers, researchers and regular extension workers, both women and men, in the different venues of the National University of Luján. The data were collected to analyze the workload of the paid and unpaid tasks carried out in their academic specialties in the different venues of the University. It is considered that the survey carried out is an exploratory attempt to highlight the difference in the use of time with a gender approach. It was considered in particular, approaches of Universities of Latin America, on the Use of Time. Regarding the measurement methods, although a probabilistic sample was initially proposed per cluster, it was decided to apply both types of probabilistic and non-probabilistic sampling. The data was collected through a cross-sectional, self-administered survey, with the assistance of surveyors who applied it at selected University sites in the second semester of the second year of the project. With respect to paid work, women showed features of underemployment (less than 35 hours) or intermediate occupation, while men in overcrowding, in relation to working more than 45 hours. However, about working days, a higher female incidence was observed, particularly on weekends. In line with other studies, gender inequalities were raised in some aspects of unpaid work. Domestic work, child care and the elderly, is often a role assigned to women in the family, which were presented in the inadequacy of services to meet basic family needs, exceeding in percentage the tasks of the home made by males. However, there has been an increase in the participation of "ellos" in the daily tasks of the home.

Keywords: Surveys, gender, use of time.

Bayesian based algorithm for the optimization of biological clocks in humans

Juan R Gonzalez[1], *Gerardo Alfonso*[2], *Dolors Pelegri*[3]

[1]juanr.gonzalez@isglobal.org, Bioinformatics Research Group in Epidemiology, Barcelona Insitute for Global Health (ISGlobal), CIBER Epidemiologia y Salud Publica (CIBERESP) and Department of Mathematics, Autonomous University of Barcelona (UAB)

[2]gerry.alfonso@swhysc.com, Department of Genetics, Autonomous University of Barcelona

[3]dpelegri@gmail.com, Bioinformatics Research Group in Epidemiology, Barcelona Insitute for Global Health

There are several evidences that DNA methylation is related to aging and, hence, overall mortality. Horvath's clock is the most commonly used biological clock based on methylation data. It uses elastic net methodology to build an age predictor which showed good prediction accuracy. We propose to use neural networks (NN) after dimensionality reduction to improve age prediction as there is no indication that the level of methylation and the chronological age of the patient should follow a linear relationship. In particular, we use neural networks trained with Bayesian learning on the CpGs obtained after dimensionality reduction. We shown that our proposed methodology outperforms the results obtained when using Horvath's method, neural networks applied to the whole methylome or when using other training algorithms like Levenberg-Marquardt. The R-squared value obtained when using our proposed approach in empirical (out-of sample) data was 0.934, compared to 0.914 using a different training algorithm (Levenberg Marquard) or 0.910 when applying the neural network directly (e.g. without reducing the dimensionality of the data first). We also demonstrate that building an age predictor using a Bayesian based algorithm after dimensionality reduction of the methylome provides accurate age predictions in three independent dataset not used during the training phase. Our approach obtained better R-squared values and root-mean-square-error than Horvath's method in those three datasets (R-squared values ranging from 0.40 to 0.70). The predictive model is implemented in an R function available through a Bioconductor package created for age prediction using methylation data encapsulated using standard Bioconductor classes (e.g. GenomicRatioSet, MethyGenoSet, ExpressionSet). As a result, we provide an accurate framework that will help to elucidate the role of DNA methylation age in complex diseases or traits related to aging.

Keywords: Bayesian networks, epigenetics, aging.

Bayesian based algorithm for the optimization of biological clocks in humans

L.F. Grajales[1], *L.A. López*[1], *O.O. Melo*[1], *R. Ospina*[2]

[1]lfgrajalesh@unal.edu.co, Departamento de Estadistica, Universidad Nacional de Colombia

[2] Departamento de Estatistica, Universidade Federal de Pernambuco, Brasil.

In this work, we propose and develop a doubly restricted exponential dispersion model, i.e. a varying dispersion generalized linear model with two sets of restrictions, a set of linear restrictions for the mean response, and at the same time, for another set of linear restrictions for the dispersion of the distribution. This model would be useful to consider several situations where it is necessary to control/analyze drug-doses, active effects in factorial experiments, mean-variance relationships, among other situations. A penalized likelihood function is proposed and developed in order to achieve the restricted parameters and to develop the inferential results. Several special cases from the literature are commented on. A simply restricted varying dispersion beta regression model is exemplified by means of real and simulated data. Satisfactory and promising results are found.

Keywords: Two-parameter exponential family, linear restrictions, penalized likelihood.

Is arsenic, cadmium and zinc exposure causally associated with renal damage? A Mendelian randomization study

Maria Grau-Pérez[1], *Jose D. Bermúdez Edo*[2], *Jose L. Gómez-Ariza*[3], *Zoraida Soriano-Gil*[4], *J.Antonio Casasnovas*[4], *Josep Redón*[1], *Maria Téllez-Plaza*[5]

[1]maria.grau.perez@gmail.com, Area of Cardiometabolic and Renal Risk, Institute for Biomedical Research INCLIVA, Valencia, Spain

[2]Department of Statistics and Operational Research, University of Valencia, Valencia, Spain

[3]University of Huelva, Huelva, Spain

[4]Fundación Instituto de Investigación Sanitaria de Aragón (IIS Aragón), Zaragoza, Spain

[5]Department of Chronic Diseases Epidemiology, National Center for Epidemiology, National Institutes for Health Carlos III, Madrid, Spain

Introduction: Traditional epidemiologic studies are limited to evaluate correlational associations between an exposure and the outcome due to confounding and other biases linked to observational data. However, Mendelian randomization (MR) is the use of genetic variation as instrumental variables to assess the causal effect of modifiable exposures on health outcomes, using observational data. Using a MR approach, we investigated the causal relationship of arsenic (As), cadmium (Cd) and zinc (Zn) exposure with renal damage among 1323 participants of the Aragon Workers Health Study (AWHS) from Spain.

Methods: Exposure to As, Cd and Zn was evaluated by its concentrations in urine. Renal damage was assessed as the urine albumin levels standardized by creatinine concentrations. We first conducted a genome-wide association analysis to identify single nucleotide polymorphisms (SNPs) associated with increased levels of urine metals in the Strong Heart Family Study, and independent database. 45 identified SNPs were further genotyped in the AWHS for this analysis. We validated 2, 2, and 1 SNPs associated with genetically elevated urine As, Cd and Zn levels, respectively. We evaluated the correlational association of As, Cd and Zn urine levels with urinary albumin using traditional linear regression models and also explored the causal association using a one-sample MR approach using the 2-stage least squares method.

Results: In traditional adjusted linear regression models, the geometric mean ratio (95% confidence interval) of urinary albumin by an interquartile range increase in As, Cd and Zn levels was 1.03 (0.97, 1.10), 1.06 (1.01, 1.11) and 1.08 (1.03, 1.15), respectively. The F-statistic of the SNPs for As, Cd and Zn were 14.8, 8.7 and 6.4, respectively, suggesting that the instrumental variables might be weak, specially for Cd and Zn. Using the MR analysis there was no evidence for a causal association of As, Cd and Zn with urinary albumin.

Conclusion: While increased urine levels of Cd and Zn were significantly associated with higher albumin levels when using a traditional association approach, the MR analysis suggested that these associations might not be causal. Mendelian randomization is an interesting tool to disentangle causal associations were randomized controlled trials are not feasible. More research is needed to confirm these findings, especially with a larger set of instrumental variables that strengthen the MR analysis.

Keywords: Mendelian randomization, metals exposure, renal damage.

Orthogonal two-way Projections to Latent Structures (O2PLS) regression in an integrative omics application

Gutiérrez Botella, Jesús[1], *Barceló Cerdá, Susana*[2]

[1]jegubo@alumni.uv.es, MSc. Biostatistics, Universitat de València

[2]sbarcelo@eio.upv.es, Department of Applied Statistics and Operational Research and Quality, Universitat Politècnica de València

Numerous methods have been developed in order to study omics datasets (proteomics, metabolomics, transcriptomics, ...), such as Principal Component Analysis or Partial Least Squares, frequently based on latent structures models. In the last years, omics datasets have become easier to obtain, and new statistical methodologies are needed to extract as much biological information as possible. These datasets often exhibit high correlation, low signal-to-noise or rank deficiency.

It is frequent to find studies in which proteomics and genomics, or transcriptomics and proteomics data are used in order to achieve biological insights. Thus, new statistical integration methodology is needed to analyse the available sources simultaneously to extract as much information as possible. A methodology based on a revision of Partial Least Squares (PLS) regression, Orthogonal 2-way Projections to Latent Structures (O2PLS), is proposed in this work, which can integrate information from two different sources.

The PLS (Partial Least Squares) regression is a multivariate method of projections in latent structures through partial least squares. This method requires two variable matrixes, X or predictive matrix, and Y or response matrix. It reduces the variables dimension of both X and Y matrixes by projecting X and Y in the space of the latent variables. These new latent variables explain the maximum variability of the variables in X, and they are as predictive as possible of the variables in Y.

However, the PLS regression has an important inconvenient. Although this regression measures the relation between both the predictive and the response matrix, there can be systematic variability inherent to either of the two matrixes and orthogonal to the other one which can difficult the model interpretation.

The O2PLS method is a revision of the PLS regression described in Trygg J, 2002. It divides the total variability of the dataset in two portions: the first one, with the non-related variability between X and Y, which will be excluded; and the second one, with the joint variation of X and Y, which will be used to estimate the final model. This O2PLS model is very useful with this kind of study, because the relations that can exist between variables in both matrixes can be contaminated by orthogonal variables in a conventional PLS, but they are removed in an O2PLS.

This methodology, applied to a metabolomics (Y) and proteomics (X) dataset, improves the interpretation of results showing relations between metabolomics and proteomics which are not present if the applied regression is a common PLS. Thus, O2PLS regression may be a reasonable alternative to conventional PLS since changes in environmental conditions can affect to the whole proteome or metabolome, but the sole purpose is to study the relations between both omics data sources.

Keywords: Omics data integration, O2PLS, latent structures modelling.

SISSREM: Shiny Interactive, Supervised and Systematic report from REpeated Measures data.

Pablo Hernández-Alonso[1], *Núria Pérez Álvarez*[2]

[1]pablo.hernandez@fimabis.org, Instituto de Investigación Biomédica de Málaga (IBIMA), Málaga University, Málaga, Andalucía, Spain; Human Nutrition Unit, Rovira i Virgili University, Reus, Catalonia, Spain.

[2]nperez@flsida.org, Department of Statistics and Operations Research, Technical University of Catalonia-Barcelona Tech. Fight against AIDS Foundation, HUGTIP.

Longitudinal methods are the procedures of choice for scientists who view their phenomena of interest as dynamic. Linear mixed models (LMM) can be used to describe relationships across time in a longitudinal dataset successfully leading with dependent observations and adding the flexibility of random effects. However, other statistical methods such as the repeated measures ANOVA can perform analysis for dependent observations, but their limitations lead to misuse by part of researchers from biomedical areas due to its statistical simplicity compared with LMM.

We have developed a Shiny R code-based application (SISSREM, Shiny Interactive, Supervised and Systematic report from REpeated Measures data) intended to be used by users in biomedical areas with low-to-medium skills in statistics. Our Shiny application is able to: i) instruct the user in the understanding of a LMM analysis for repeated measures with an example database; ii) allow the user to analyse their own data; and iii) allow the user to create an interactive, supervised and systematic Rmarkdown report to be exported from the Shiny application. The main core of the application consists of a guided walk through a default analysis with an example database and the systematic decisions that must be performed in an LMM analysis. Therefore, we have structured the application into three main modules according to their application: i) exploratory data analysis (EDA) to gain insight into data; ii) graphic module to check the relationship between the variables of interest; iii) fitting module to perform the LMM together with evaluating significance of the constructed LMM (e.g. likelihood ratio test). This application will be published online by the end of June. Importantly, its code will be accessible in order to be updated or adapted for other purposes. SISSREM is a functional Shiny application which is intended to spread the usefulness of LMM into the biomedical research area.

Keywords: SISSREM, linear-mixed model, longitudinal data,shiny app.

AMS: 62M10

Modeling the mortality risk of a disease from total mortality

Lidia Herrero Huertas[1], *Paloma Botella Rocamora*[2]

[1]lidia.herrero.huertas@gmail.com, Servicio de Estudio Epidemiológicos y Estadísticas Sanitarias, Dirección General de Salud Pública, Generalitat Valenciana.

[2]botella_pal@gva.es, Servicio de Estudio Epidemiológicos y Estadísticas Sanitarias, Dirección General de Salud Pública, Generalitat Valenciana.

Exploring the mortality of a territory is very useful to describe, analyze and understand the health status of a population. In the literature, mortality can be seen from two perspectives: a global view that studies mortality from all cause of death or another more specific view that studies mortality of each disease. This work considers both approaches at the same time taking advantage of the behavior of the risks of total mortality to explain, in part, the risks of dying from a disease. The objective is to extract the differential behavior between both mortalities. Thus, the peculiarities of each disease that differ from the behavior of the general mortality of the population will be detected.

About the methodology, it is framed within Bayesian spatial modeling. Particularly, multivariate hierarchical models have been applied and adapted in this work. This approach has been made possible to estimate and represent the standardized mortality ratios (SMRs) of a disease, incorporating the estimate of the total risk that could be expected for all causes of death in a given area. This proposal allows to represent on a map three types of relative risks: (1) those associated with total mortality, (2) those linked to the disease itself and (3) those that show the differential behavior between both mortalities. In this way, if an area has a significant risk, the vision presented by this work will show its origin; either it will be part of the general behavior of mortality or it will stand out in the specific risk pattern of that disease. On the other hand, this methodology allows to extract the correlations between the patterns of both mortalities, total and specific. It will indicate that the general behavior appears in the pattern of many of the studied diseases.

These models are applied for a mortality study of Comunitat Valenciana in the period between 2011 and 2015. Significant differences are detected between men and women in each of ten selected most important causes of death. This selection of diseases is based on frequency of observed cases or their importance in the field of epidemiology.

In short, the decomposition of the mortality of a disease exhibiting its general and specific behavior, is a very useful tool that provides more detailed information on the health of the population of each territory. In particular, this study shows different behaviors between sex. This allows to develop strategies to transform this situation and achieve the equity gender.

Keywords: Mortality, risk, pattern, behavior.

Dealing with missing predictor variables in logistic regression models with complex survey data

Amaia Iparragirre[1], *Irantzu Barrio*[2], *Jorge Aramendi*[3], *Inmaculada Arostegui*[4]

[1]amaia.iparragirre@ehu.eus, Departamento de Matemática Aplicada, Estadística e IO, UPV/EHU
[2]irantzu.barrio@ehu.eus, Departamento de Matemática Aplicada, Estadística e IO, UPV/EHU
[3]j-aramendi@eustat.eus, Eustat - Euskal Estatistika Erakundea - Instituto Vasco de Estadística
[4]inmaculada.arostegui@ehu.eus, Departamento de Matemática Aplicada, Estadística e IO, UPV/EHU.
BCAM - Basque Center for Applied Mathematics

Researchers in social and health sciences are increasingly interested in using complex survey data. One of the main characteristics of the data collected by a complex sampling design is the presence of the sampling weights, which are the number of individuals that one observation represents in the population. However, there is still a lack of statistical methodology developed for this type of data. The first discussion in this context has been the importance of incorporating the sampling weights in order to obtain an accurate estimation of the prediction model. Specifically, for a dichotomous response variable and in the context of a logistic regression model, the pseudo-likelihood function has been proposed as a modified version of the likelihood function. Nevertheless, there are many questions to be considered in the prediction framework (selection of the predictors, the functional relation between outcome and predictors, imputation of missing values, evaluation of calibration and discrimination ability, among others) which are relevant to end up with a good validated prediction model. Therefore, we believe it is necessary to consider the design of the survey, in particular the weights, throughout the process of development and validation of a prediction model, beyond the estimation of its parameters. In this work we study the impact the sampling weights have in the imputation process of a relevant missing predictor variable and we propose a new approach to impute a dichotomous predictor variable in practice.

This methodology has been applied to the ESIE survey data, which have been designed and collected by the Official Statistics Basque Office (Eustat) in order to estimate aspects related to the use of technology of companies in the Basque Country. Previous year's response has been seen to be a good predictor, but this information was available only for those companies that were sampled the year before. Therefore, we applied the proposed methodology to these data improving considerably the goodness-of-fit, the prediction and the discrimination ability of the model.

Keywords: Complex survey data, imputation, cutpoints.

Development and evaluation of a sequential adaptative sampling strategy to delimiting the distribution of *Xylella fastidiosa*: a case study in Alicante

E. Lázaro[1], *D. Conesa*[2], *A. López-Quilez*[2], *V. Dalmau*[3], *A. Ferrer*[3] *and A. Vicent*[1]

[1]lazaro_ele@gva.es, Centre de Protecció Vegetal i Biotecnologia, Institut Valencià d' Investigacions Agràries, Moncada

[2] Departament d' Estadística e Investigació Operativa, Universitat de València, Burjassot

[3] Servei de Sanitat Vegetal, Conselleria d'Agricultura, Medi Ambient, Canvi Climàtic i Desenvolupament Rural, Silla

Xylella fastidiosa is a phytopathogenic bacterium regulated in the European Union (EU) to avoid its introduction and spread within all Member States. The current legal provisions specify the implementation of an intensive surveillance progam in those regions in which the presence of disease was confirmed. The main aim of this plan is to make an accurate delimitation of the geographic extent of the disease to further implement eradication measures. Alicante, is one of the EU regions affected by *X. fastidiosa*. As a consequence, the area is being subjected to surveillance and sampling actions. Specifically, since the disease was first detected in June 2017 approximately 101,300 has. have been surveyed and around 20,000 samples have been taken and analysed for *X. fastidiosa*. These actions imply a great economic investments, thus can we help risk managers to decide in which areas invest greater effort in surveillance and sampling?, how many samples are necessary to achieve a reasonably accurate delimitation of the extent of the disease?.

Based on 2018 official survey data, different sampling and sampling-surveillance strategies were compared aiming to improve effectiveness. Sampling strategy is based on limiting the number of samples according different spatial resolutions. We implement an algorithm to optimise the cutoff number of samples by simulating differet random sampling scenarios from the reference data. Sampling-surveillance strategy is based on tailoring surveillance and sampling intensity by combining an adaptative approach. The adaptive approach has the purpose of improving the accuracy of the delimitation by exploiting the typical spatial aggregation of *X. fastidiosa*. We suggest a three-phase design in which surveillance and sampling efforts are adaptively allocated in those spatial units where disease has been detected in the previous phase. We implement an algorithm to optimise the number of spatial units to be surveyed and the sampling intensity in each step by simulating different random sampling scenarios from the reference data. Evaluations are quantified comparing the delimitation efficacy and disease prevalence estimates between the proposed strategies and the reference data.

Keywords: Spatial sampling, adaptive sampling, simulation-optimization.

Predicting rainfall soil erosivity and soil properties in the Basque Country with Geoadditive models

Dae-Jin Lee[1], *Lore Zumeta Olaskoaga*[1], *María Xosé Rodríguez Álvarez*[3], *Ander Arias*[4], *Nahia Gartzia*[4], *Ainara Artetxe*[4]

[1]dlee@bcamath.org, [1]lzumeta@bcamath.org, BCAM - Basque Center for Applied Mathematics,
[3]mxrodriguez@bcamath.org, Ikerbasque & BCAM - Basque Center for Applied Mathematics,
[4]agonzalez@neiker.eus, [4]ngartzia@neiker.eus, [4]ahiartetxe@neiker.eus, Neiker-Tecnalia Basque Institute for Agricultural Research and Development

Rainfall is one of the main drivers of soil erosion and land degradation which affects landscape and human activities (e.g., agricultural productivity and forestry management). Indeed, soil erosion depends on several factors such as rainfall distribution (e.g., intensity, duration, magnitude, cumulative per event), temperature, elevation, longitude, latitude and it is controlled by the interactions between lithology, orography, hydrography, land use, and vegetation. One of the most commonly used soil erosion models is the Universal Soil Loss Equation (USLE) and its revised version (R)USLE which calculates the average rainfall erosivity or simply the R-factor. In this work, we proposed Geoadditive models as a general framework for predicting soil erosivity in the Basque Country based on several covariates, including monthly variation to account for seasonality. We also extend the methodology to the analysis of soil texture, organic matter and carbon stock to provide maps that will contribute to a more efficient management and decision-making process for the agricultural sector.

Keywords: Spatial prediction, Generalized Additive Models, soil erosion.

Modeling microbiota time-series data

J. López[1], *F.J. Santonja*[2]

[1]jesua@alumni.uv.es

Departament d'Estadística i Investigació Operativa. Universitat de València

[2]francisco.santonja@uv.es

Departament d'Estadística i Investigació Operativa. Universitat de València.

The human microbiota is the collective genomes of the microorganisms that live inside and on the human body. It is composed by many types of microbes as bacteria, archaea, fungi, protists and viruses. These microorganisms resides either in or on of a number of human tissues and biofluids, including the skin, mammary glands, placenta, seminal fluid, uterus, ovarian follicles, lung, saliva, oral mucosa, conjunctiva, biliary and gastrointestinal tracts. Microbiota have been found to be crucial for the health of their host. The human microbiota and their interactions with the host play an important role in basic biological processes and in the development and progression of major human diseases such as infectious diseases, liver diseases, gastrointestinal cancers, metabolic diseases, respiratory diseases, mental or psychological diseases, and autoimmune diseases.

Note that microbiota is inherently dynamics. Thus, longitudinal microbiota analysis can provide rich information of short-and long-term trends of microbial communities. In addition, a common way to express the microbiota data is using relative proportions. As a consequence, compositional data analysis is a valid aproach to analyze the microbiota sequencing data.

In this work, we propose a dynamical model in order to analyze trends of microbiota time-series with a Dirichlet response variable. We will present different modeling proposals and estimation methods.

Keywords: Microbiota, multivariate time series, dynamic models, Dirichlet.

Understanding the gender gap in STEM careers. A longitudinal data analysis

Emilia López-Iñesta[1], *Anabel Forte*[2], *Silvia Rueda*[3], *Carmen Botella*[3], *Paula Marzal*[4]

[1]emilia.lopez@uv.es, Departamento de didáctica de la matemática, Universitat de València

[2] anabel.forte@uv.es Departamento de Estadística e I.O., Universitat de València

[3] Departamento de Informática, Universitat de València

[4] Departamento de Ingeniería Química, Universitat de València

In recent years it is increasingly evident that, among others, health studies and policies have an important gender gap. Understand why and remedy it becomes, then, an important challenge that starts by looking at who studies what.

Recent studies show that the number of female students enrolled in STEM related disciplines (such as Data Sciences) have been decreasing in the last twenty years, while the number of women resigning from technological job positions remains unacceptably high. In this paper, we try to show the effects of a working program developed by the School of Engineering at the University of Valencia (ETSE-UV), Spain, which aims at decreasing the gender diversity gap as well as increasing and retaining the number of female students enrolled in STEM fields. The data analysis so far, establishes that, in part, this program has helped to achieve higher female graduation rates, especially among Bachelor students, as well as increasing the number of top-decision positions held by faculty women.

The goal of this work is to use models for longitudinal data to understand if the temporal evolution of the gender gap shows differences related to the specific area of STEM as well as to understand the most effective key points in the STEM careers promotion policies.

Keywords: Longitudinal data, woman in STEM.

Multivariate disease mapping for crime-related outcomes

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

[1]antonio.lopez@uv.es, Department of Statistics and Operational Research, University of Valencia

[2] Department of Social Psychology, University of Valencia

Bayesian spatial modeling is increasingly used for studying different social problems, including crime and violence. However, there is less research that has analyzed crime-related outcomes from a multivariate perspective. The aim of this study is to analyze the shared spatial structure for three types of police calls.

The study was conducted in the city of Valencia (Spain). Valencia Police Department provided information about the calls received to the call service 092 from 2010 to 2016. All police calls were aggregated at the census block group level. Three types of police calls were selected for this study: Alcohol-related (N = 14,570); vandalism (N = 7,346); and fights (N = 26,624).

First, different Bayesian Poisson spatial regression models were conducted for each outcome (i.e., the number of police calls). Second, we conducted a spatial principal component analysis, which allows to assess a multivariate joint distribution for a large number of outcomes. Specifically, we used the M-based modeling.

Results showed a high correlation between the relative risks of the spatial univariate models for alcohol, vandalism and fights-related police calls. Mapping the log relative risk for each outcome we observe, in general, a higher relative risk in the center of the city and in the eastern areas. This pattern is especially pronounced for alcohol-related police calls.

In addition, the multivariate approach provided the eigenvectors for three components. We obtained a eigenvalue of 2.75 for the fist component, 0.18 for the second component, and 0.07 for the third component. This result indicates that the fist component is explaining almost 92% of the spatial variability. This first component points a spatial distribution where the three police calls types are equally contributing. The map of the shared spatial distribution shows a clear pattern where there is a higher risk of police calls in the center, some areas in the west, and the northeast of the city, independently of the type of call.

Keywords: Bayesian inference, multivariate model, crime mapping.

Role of aspirin in primary prevention of cardiovascular disease in real world data. A propensity score match.

Cristina López Zumel[1], *Josep Redón i Mas*[2], *Jose Luis Holgado Sánchez*[3], *Inma Saurí Ferrer*[4], *Antonio Fernández Giménez*[5], *Adrián Ruíz-Hernández*[6], *Fernando Martínez García*[7]

[1]crislozu@incliva.es, Incliva
[2]josep.redon@ uv.es, Incliva
[3]jlholgado@incliva.es, Incliva
[4]isauri@incliva.es, Incliva
[5]afernandez@incliva.es, Incliva
[6]anruhe@gmail.com, Incliva
[7]fernandoctor@hotmail.com, Incliva

The use of low-doses of aspirin in secondary prevention is well documented although its use is associated with an increase of bleedings, especially of the gastrointestinal tract. Although the controversy still remains, there is now compelling evidence against the use of aspirin in primary prevention. Therefore, our aim was to evaluate the effects of long-term low-doses of aspirin use in the primary prevention of cardiovascular events as well as the risk of major bleeding using real world data.

Electronic Health Records (EHR) of the whole Valencia Community from 2012 were used for this study, which is a part of Bigmedilytics project (Big Data for Medical Analytics) financed by the European Commision. Data consists on a total of 3.066.449 patients aged 18 and over: 44.158 (1,5%) were taking low-dose aspirin in primary prevention and 3.022.291 (98,5%) were not. Different databases were used such Abucasis which contains the ambulatory information, GAIA with information about prescriptions and Orion with information regarding to hospitalizations and mortality. As the major cardiovascular outcomes we considered coronary heart disease, ischemic stroke, peripheral artery disease and all cause mortality. As potential adverse effects for the aspirin use we considered hospitalizations for bleeding events and the development of ferropenic anaemia. We excluded the patients that suffered any of the events before the inclusion date. We performed a nearest neighbor matching technique based on logit regression to obtain a control group with similar aspirin-group cardiovascular risk. A total of 17.896 individuals were included in the study: 50% in control group, 50% in aspirin group. For the matching we adjusted by age, gender, estimated glomerular filtration rate by CKD-EPI, type 2 diabetes, hypertension, dyslipidemia and tobacco use. Cox proportional hazards analysis and Kaplan-Meier survival analysis are used to assess differences in mortality and cardiovascular events.

Our real world data results indicate that low-doses long-term aspirin does not prevent the incidence of cardiovascular events in primary prevention and may be a trend of increase the bleeding risk.

Keywords: Aspirin, propensity score, cardiovascular disease.

A Bayesian regression model for the non-standardized t distribution with location, scale and degrees of freedom parameters

Margarita Marin Jaramillo[1], *Edilberto Cepeda Cuervo*[2]

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

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

In this work we analyse situations where the variable of interest cannot be assumed to has normal distribution, mainly because of the presence of heavy tails. Although from a Bayesian perspective. This problem has been widely analysed (Geweke, 1993; Fernández & Steel, 1998; Gelman et al., 2006; Fonseca, Ferreira & Migon, 2008), none of these works develop joint regression modelling of location and scale parameters. Thus, we propose a Bayesian t regression models in which the mean and the scale parameters follows regression structures.

To fit the proposed model, we develop an extension of the Bayesian method proposed by Cepeda (2001) and Cepeda & Gamerman (2005) to obtain samples of the conditional posterior distribution of the regression parameters and degrees of freedom. Thus, we propose a Bayesian method that includes a new proposal to obtain samples of the posterior conditional distribution of the degrees of freedom.

Finally, we apply the proposed Bayesian t regression models to analyse of simulated and real dataset.

Keywords: Non-standardized t model, Bayesian regression.

Bibliography

Cepeda, C. (2001), 'Modelagem da variabilidade em modelos lineares generalizados', Unpublished Ph. D. tesis. Instituto de Matemáticas. Universidade Federal do Rio do Janeiro. ////http://www. docentes. unal. edu. co/ecepedac/docs/M http://www. bdigital. unal. edu.co/93942.

Cepeda, E. & Gamerman, D. (2005), 'Bayesian methodology for modeling parameters in the two parameter exponential family', Revista Estadística, 57(168-169), 93-105.

Fernández, C. & Steel, M. F. (1998), 'On Bayesian modelling of fat tails and skewness', Journal of the American Statistical Association, 93(441), 359-371.

Fonseca, T. C., Ferreira, M. A. & Migon, H. S. (2008), 'Objective Bayesian analysis for the student-t regression model', Biometrika, 95(2), 325-333.

Gelman, A. et al. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)', Bayesian analysis, 1(3), 515-534.

Geweke, J. (1993), 'Bayesian treatment of the independent student t linear model', Journal of applied econometrics, 8(S1).

Longitudinal zero-inflated regression model for missing count data

Oscar Orlando Melo Martínez[1], *Danny Samuel Martínez Lobo*[2], *Sandra Esperanza Melo Martínez*[3]

[1]oomelom@unal.edu.co, Department of Statistics, Universidad Nacional de Colombia
[2]dsmartinez@unbosque.edu.co, Department of Psychology, Universidad del Bosque
[3]semelom@unal.edu.co, Department of Agronomy, Universidad Nacional de Colombia

Longitudinal zero-inflated count data arise frequently in research when assessing the effects of number of bees returning to the hive with loads of pollen and the efficiency of genetically modified corn in relation to conventional corn in the control of Spodoptera frugiperda. In these applications, the distributions inflated by zeros are a good alternative and the longitudinal models with variable response inflated by zeros with missing information is no stranger in the previous two applications. Therefore, we propose a methodology for the estimation and imputation of the missing information in the longitudinal response variable with Poisson or Binomial Negative distributions inflated by zeros. The data is supposed to be missing at random (MAR), and in each time, the algorithm EM is used. In step E, a weighted regression is carried out, conditioned to the previous time that is taken as covariables. In step M, the estimation and imputation of the missing data is carried out following the methodology proposed.

In the two applications, the complete data was taken and a random loss of 20%, 30%, 40% and 50% was carried out. The estimated values had average successes of 72% with respect to the original data and the confidence intervals of the simulations contain the estimated parameters of the model made with the corn study data (Zero-Inflated Poisson). On the other hand, the estimated values had an average of 51.6% with respect to the original data, and in three of the four simulated scenarios, the confidence intervals of the simulations contain the estimated parameters of the model made with pollen study data (Zero-Inflated Negative Binomial).

Keywords: Longitudinal zero-inflation model, count data, missing-data imputation.

AMS: 62L12, 62H86, 91G70

Zero-inflated models for regression analysis of count data applied to an experiment in agronomy

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, Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas

This paper presents how to perform data modeling of experiments in Agronomy or Biology when there is an excess number of zero counts and also over-dispersion in experiments where the response variable usually is a count data. This study corresponds to a block design; it was conducted in the municipality of Fómeque Cundinamarca Colombia with two treatments: fertilization plans of 6 and 8 kg per plant whose response variable is flowers with botrytis in tomato plants. Well-nourished plants are expected to be more tolerant to disease.

Hurdle and zero-inflated models were developed to accommodate both the excess of zeros and skewness of the data, covariate effects can be incorporated into both components of the models. The Hurdle model is a two-component mixture model consisting of zero and the non-zero observations component following a conventional count distribution. Modeling the count component as a negative binomial distribution is significantly higher than a Poisson distribution which not solves properly the problem studied.

The negative binomial zero-inflated models provide a better fit compared to Hurdle models in the application used. However, experiments in agronomy usually collect samples longitudinally, which introduces time-dependent and correlation structures among the samples and thus further complicates the analysis and interpretation of the data. Having measurements in time is not the most appropriate because the correlation of plants at different times is being ignored.

Therefore, the negative binomial zero-inflated longitudinal model that models the correlation and gives a better fit to the studied data was also considered. In this paper, we propose to use negative binomial mixed models (NBMMs) for longitudinal data in Agronomy; this methodology was proposed by Zhang (2018). The proposed NBMMs can efficiently handle over-dispersion very frequently in that kind of experiments and varying total reads, and can account for the dynamic trend and correlation among longitudinal samples.

The extended NBMMs can include various types of fixed effects and random effects, and can incorporate various correlation structures among observations within the same subjects. This methodology develops an efficient and stable IWLS (iterative weighted least squares) algorithm to fit those extended NBMMs by taking advantage of the standard procedure for fitting linear mixed models. When were compared between those models NBMMs methodology is the better a good alternative for data analysis in our application.

Keywords: Hurdle model, zero-inflated Negative Binomial model, over-dispersion.

Spatial conditional overdispersion models. Modelling infant mortality rates

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

[1]mabel.morales@ehu.eus, Departamento de Economía Aplicada III (Econometría y Estadística), Universidad del País Vasco UPV/EHU, Bilbao

[2]vicente.nunezanton@ehu.eus, Departamento de Econometría y Estadística (E.A. III), Universidad del País Vasco UPV/EHU, Bilbao

Regression models for count data often present overdispersion, a phenomenon that arises when the real variance of the data is larger than the one specified in the model. In addition, when working with georeferenced data, the spatial dependence that may exist among the different locations must be taken into account in order to produce reliable inference processes from the estimations. In this work, we revise generalized spatial conditional overdispersion models for count data, where spatial neighborhood structures for the mean and for the overdispersion parameter are specified. This allows researchers to be able to quantify the spatial association that might exist in these structures, and also to model the possible existing overdispersion, which is assumed to be generated by these spatial relations. We analyze infant mortality rates from the different departments in Colombia, and illustrate the usefulness of the different spatial conditional overdispersion models when analyzing this type of data. Additionally, we also assess the performance of such models when including a conditionally autoregressive (CAR) spatial random effect in the model specification. Models have been fitted with the use of the Markov Chain Monte Carlo (MCMC) algorithms within the context of Bayesian estimation methods.

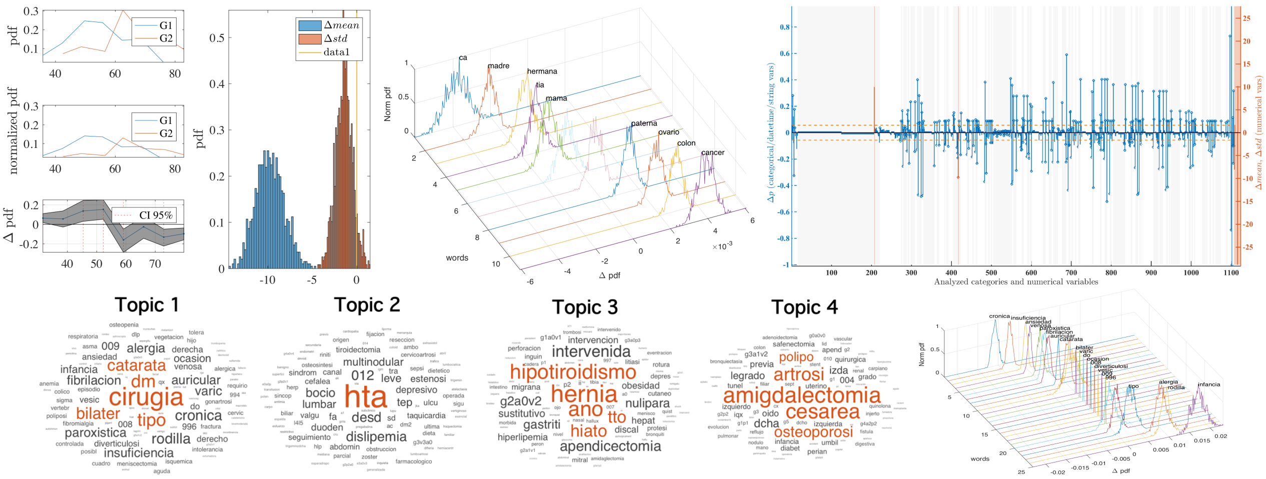
Keywords: Overdispersion models, spatial models, Bayesian methods.

Distribution-test estimation and topic modeling from Big-Data Analytics from heterogeneous hospital records in ovarian cancer

S. Muñoz-Romero, J. García-Donas, E. Sevillano, L. Bote-Curiel, M. Yagüe, N. Lainez, E.M. Guerra, M. Garrido, T. García-Donas, S. Amarilla, P. Navarro, S. Ruiz, M.D. Fenor, J.F. Rodríguez-Moreno, J.L. Rojo-Álvarez

Department of Signal Theory and Communications, Universidad Rey Juan Carlos
Fundación Hospital de Madrid
Clara Campal Comprehensive Cancer Center, Madrid
Complejo Hospitalario de Navarra
Hospital Ramón y Cajal, Madrid
Hospital de Leganés.

Prognosis of advanced ovarian cancer is dismal with most cases recurring after initial surgery. We explored the potential of Big-Data Analytics (BDA) to screen for clinically relevant variables in currently available hospitalary departamental information systems. An observational study with two cohorts (one prospective and one retrospective) was designed. Inclusion criteria were adult patients (>18 years old) diagnosed with epithelial ovarian cancer stage IC or superior. Clinical and histological data were recorded by a central data manager to ensure homogeneity in data collection. BDA built approximations to the statistical distribution of the tests for different feature types (metric, categorical, and free text). Up to 265 patients in four different hospitals were recruited. Median age was 59 years (range 20-87), stage distribution was 48 (18%) I, 20 (8%) II, 122 (46%) III, 41 (15%) IV, and 34 (13%) NA. The distributions characterized the confidence for average differences and distribution profiles (up, left), proportion differences (up, middle), text-profile differences, and an a database-landscape view (up, right), for exitus vs non-exitus grouping. Free-text variables in medical history were studied by Latent Dirichlet Allocation, obtaining 4 topics in terms of word clouds (down,left) and their distributions were also estimated (down, right). The proposed system identified several variables of interest, e.g., a higher frequency of upfront surgery (vs interval surgery) and bevacizumab administration (vs chemotherapy alone) in the non-exitus group. Simple Big-Data analytics can help to identify new prognostic factors for clinical practice.



Keywords: Bootstrap resampling, big data analytics, oncology.

Multi environment trial analysis to determine the resistance of cereal varieties to *Pratylenchus neglectus* in Australia

Isabel Muñoz Santa[1], *Joshua Fanning*[2], *Katherine Linsell*[3]

[1]isabel.munoz-santa@uv.es, Department of Statistics, University of Valencia

[2]joshua.fanning@ecodev.vic.gov.au, Department of Economic Development, Job, Transport and Resources, Agriculture Victoria

[3]katherine.linsell@sa.gov.au, South Australia Research and Development Institute

Pratylenchus neglectus (*P. neglectus*) are root lesion nematodes which reduce plant production. They are considered a pest for cereals crops in Australia and reported to cause major yield losses with a high economic cost to the grain industry. The recommended approach to control the density of these nematodes in the soil is by breeding varieties which are resistant; i.e. varieties which are able to inhibit the reproduction of these nematodes in the field.

In this study, we conducted 11 field trials from 2011 to 2017 in South Australia and Victoria with the objective of providing resistance ratings of 89 different cereal varieties under high and low levels of pre-established nematode densities in the field. Multi environment trial analysis were used to analyse the data where the term environment refers to each year by location by nematode density combination. Spatial techniques were used to account for the spatial variability in each trial and a factor analytic model was fitted to model the genotype by environment interaction effects. This approach allowed us to predict nematode DNA levels of each variety at each environment, asses the genetic correlation between pairs of environments and check for consensus of resistance ratings for the varieties across all environments.

Multi environment trial analyses revealed high genetic correlation between all pairs of environments which is related to consistent performance of varieties across environments. Therefore, varieties can be rated according to their resistance across a wide area and over different year conditions. This provides accurate and valuable information to growers in relation to the resistance ratings of the varieties evaluated.

Keywords: Resistance to Pratylenchus, multi-environment trial analysis, factor analytic models.

DNA methylation signatures and incident cardiovascular disease in the Strong Heart Study

Ana Navas-Acien[1], *Arce Domingo-Relloso*[2], *Lizbeth Gomez*[3], *Maria Tellez-Plaza*[4], *Karin Haack*[5], *Daniele Fallin*[6], *Shelley Cole*[7]

[1]an2737@cumc.columbia.edu, Columbia University Mailman School of Public Health

[2]ad3531@cumc.columbia.edu, Columbia University Mailman School of Public Health

[3] lg2982@cumc.columbia.edu, Columbia University Mailman School of Public Health

[4] m.tellez@isciii.es, Centro Nacional de Epidemiología

[5] khaack@txbiomed.org, Texas Biomedical Research Institute

[6] dfallin@jhu.edu, Johns Hopkins Bloomberg School of Public Health

[7] scole@txbiomed.org, Texas Biomedical Research Institute

Cardiovascular disease (CVD) is the leading cause of death in the world. Several studies have highlighted the promising role of epigenetic modifications (in particular, DNA methylation) in risk prediction and early detection of disease. The objectives of this study were to investigate the association of blood DNA methylation with incident CVD, CHD and stroke in participants from the Strong Heart Study as well as to evaluate the potential ability of DNA methylation status to predict CHD beyond traditional risk factors.

We conducted an Epigenome-Wide Association Study of 2325 American Indian men and women 45-74 years old who participated in the Strong Heart Study in 1989-1991. DNAm was measured in 790026 loci with the Illumina Infinium Human MethylationEPIC platform, pre-processed and corrected for batch effect and cell heterogeneity. Cox regression was used for survival analysis, and PCA and complete-linkage clustering were used for grouping the most significant CpG sites. Area under the receiver operating characteristic curve (AUC) was used to assess the predictive ability of DNA methylation.

After correcting for multiple comparisons, we found 5 significant positions for CVD, 13 for CHD and 2 for stroke, with no overlap between them. Between the significant CpGs for CHD, we found three different clusters defined by the strength of the correlations between CpGs. The first two PCs explained 72 % of the variation in DNAm for those 13 CpGs. The AUC changed from 0.695 to 0.729 when using the first two PCs as predictors, supporting that DNAm improves the prediction ability of the general prediction equation for CHD in American Indians. These results need to be replicated in an external population.

Keywords: DNA methylation, cardiovascular disease, American Indians.

Error-rate estimation in discriminant analysis of non-linear longitudinal data: a comparison of resampling methods

Vicente Núñez-Antón[1], *Rolando de la Cruz*[2], *Claudio Fuentes*[3], *Cristian Meza*[4]

[1] vicente@nunezanton@ehu.eus, Departamento de Econometría y Estadística (E.A. III), Universidad del País Vasco UPV/EHU, Bilbao

[2] rolando.delacruz@uai.cl, Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Santiago, Chile

[3] fuentesc@stat.oregonstate.edu, Department of Statistics, Oregon State University, Corvallis, Oregon, USA

[4] cristian.meza@uv.cl, CIMFAV, Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso, Chile

Consider longitudinal observations across different subjects such that the underlying distribution is determined by a non-linear mixed-effects model. In this context, we look at the misclassification error rate for allocating future subjects using cross-validation, bootstrap algorithms (parametric bootstrap, leave-one-out, .632 and .632+), and bootstrap cross-validation (which combines the first two approaches), and conduct a numerical study to compare the performance of the different methods. The simulation and comparisons in this study are motivated by real observations from a pregnancy study in which one of the main objectives is to predict normal versus abnormal pregnancy outcomes based on information gathered at early stages. Since in this type of studies it is not uncommon to have insufficient data to simultaneously solve the classification problem and estimate the misclassification error rate, we put special attention to situations when only a small sample size is available. We discuss how the misclassification error rate estimates may be affected by the sample size in terms of variability and bias, and examine conditions under which the misclassification error rate estimates perform reasonably well.

Keywords: Parametric bootstrap, longitudinal data, non-linear mixed-effects models.

AMS: 62F40, 62J12.

Multivariate Bayesian areal data models using R-INLA: MuBAMo R-Package

Francisco Palmí-Perales[1], *Virgilio Gómez-Rubio*[2], *Miguel A. Martinez-Beneito*[3]

[1]Francisco.Palmi@uclm.com, Department of Mathematics, University of Castilla-La Mancha
[2]Virgilio.Gomez@uclm.com, Department of Mathematics, University of Castilla-La Mancha
[3]martinez_ mig@gva.es, Centro Superior de Investigación en Salud Pública, CSISP-FISABIO

The high computation time required to fit spatial models is a well known handicap of spatial statistics. Different ideas have been developed to speed up fitting spatial models. On the one hand, some authors have reparameterized the model in order to avoid high computational-time operations within the framework of MCMC algorithms (Botella-Rocamora et. al. 2013). On the other hand, some authors prefer aproximate methods, such as the INLA method which is commonly used in order to avoid the high computation time of the MCMC models. Following this last idea, in this work, a group of different multivariate Bayesian spatial models for areal data, which were proposed for been computed using MCMC algorithms and then reparametrized in order to be more computationaly efficient, have been implemented using R-INLA. Furthermore, a comparison between the different implementations has been carried out and an R package called MuBAMo (Multivariate Bayesian Areal data Models) has been created putting together all the implemented models together with other supplements, such as a simulated dataset to test the proposed models.

Keywords: Spatial areal models, R-INLA, Bayesian statistics.

AMS: 62-04

Bibliography

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

On the need of smoothing life expectancies in small areas studies

Jordi Pérez-Panadés[1], *Paloma Botella-Rocamora*[2], *Miguel Ángel Martínez-Beneito*[3]

[1]perez_jorpan@gva.es, Subd. Gral. Epidemiología, Vigilancia de la Salud y Sanidad Ambiental. DGSP. Conselleria de Sanitat Universal i Salut Pública. Generalitat Valenciana

[2]botella_pal@gva.es, Subd. Gral. Epidemiología, Vigilancia de la Salud y Sanidad Ambiental. DGSP. Conselleria de Sanitat Universal i Salut Pública. Generalitat Valenciana

[3]martinez_mig@gva.es, Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO-Salud Pública). Generalitat Valenciana

Life expectancy at birth (simply LE onwards), is the most widely used synthetic indicator characterizing the life, health, education and other social dimensions of a country or territory. Currently, technological advances and the development of georeferencing tools, allows having mortality and population data at the highest level of disaggregation. This fact, jointly with the need of providing indicators elaborated at the smallest administrative units available, makes LEs to be widely used in small areas studies.

However, the construction of this indicator on small area (in population terms) requires a rigorous analysis. For developing this indicator it is necessary to know the age-specific mortality rates for each administrative unit. These rates are derived from the age-specific distribution of mortality on those units. Thus, LEs can be only calculated in a reliably manner when the age-specific mortality rates in each unit of study are accurately known. However, when working with small areas, there are often few deaths in most of the age groups considered, what increases the uncertainty of rates and makes it harder to obtain reasonable age-specific estimates.

In this work, a multivariate random effects model has been applied to produce improved age-specific mortality rates and smoothed LEs, in small areas studies. This model takes into account dependence between consecutive age groups and spatial dependence between adjacent regions. It has been used to study the spatial distribution of smoothed LE for both sexes (separately) at different level of disaggregation: municipalities in Comunitat Valenciana and census tracts within cities (Alicante and Valencia). Differences between raw and smoothed LEs, and their relation with the size of the corresponding population, are analysed.

Keywords: Life expectancy, Bayesian hierarchical modelling, disease mapping, small areas studies.

Machine learning methods for assessing and predicting low muscle quantity and/or quality in HIV infected individuals

Nuria Perez-Alvarez[1], *Esteban Vegas*[2], *Carla Estany*[3], *Anna Bonjoch*[4], *Eugenia Negredo*[5]

[1]nperez@flsida.org, Department of Statistics and Operations Research, Technical University of Catalonia-Barcelona Tech. Fight against AIDS Foundation, HUGTIP

[2]evegas@ub.edu, Department of Genetics, Microbiology and Statistics, University of Barcelona

[3]cestany@flsida.org, Fight against AIDS Foundation, HUGTIP

[4]abonjoch@flsida.org, Fight against AIDS Foundation, AIDS Care Unit, Infectious Diseases Service, HUGTIP Universitat Autònoma de Barcelona

[5]enegredo@flsida.org, Fight against AIDS Foundation, AIDS Care Unit, Infectious Diseases Service, HUGTIP. Universitat Autònoma de Barcelona. Universitat de Vic-Universitat Central de Catalunya

The age-associated loss of skeletal muscle mass as well as diminished muscle strength and/or physical performance, is associated with reduced physical capability, impaired cardiopulmonary performance, disability, and mortality among older people. In the case of the HIV infected patients for whom the aging process starts earlier, the assessment and care of the body composition balance has become a challenge to healthy aging. We aimed to determine which variables (among demographic, fat and bone measurements) and their cut-off values are more valuable to predict low quantity or quality muscle mass, defined as Appendicular Lean Mass Index/height^2 <7.0 kg/m^2 for men and <6.0 kg/m^2 for women. This study was a crosssectional analysis which included 1475 outpatients whose mean (SD) age was 51 (10) years (60% aged $\geq$50), and 24% were female.

The available variables (87) were demographical and Dual Energy X-ray Absorptiometry (DEXA) scan values assessing fat and lean mass and bone mass density. The methods used were scatterplots and calculations of correlation coefficients for the concordance assessment and multivariate analysis to identify outliers and to determine the profile of patients with abnormal muscle mass composition. Non-supervised and supervised multivariate techniques, such as principal component analysis (PCA) and random forests for classification, were applied. A combination of variables that contain information from the data set and the cut-off values to classify the patients according to the muscle status were identified.

Keywords: Machine learning, body composition, muscle mass.

AMS: Computational methods.

Multidimensional adaptive penalised spline models: application to neurons' activity studies

María Xosé Rodríguez - Álvarez[1,2], *Maria Durban*[3], *Dae-Jin Lee*[1], *Paul H.C. Eilers*[4], *Francisco Gonzalez*[5]

[1]mxrodriguez@bcamath.org, BCAM - Basque Center for Applied Mathematics, Bilbao, Spain

[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

[3] Department of Statistics, University Carlos III of Madrid, Leganés, Spain

[4] Erasmus University Medical Center, Rotterdam, The Netherlands

[5]Department of Surgery and CIMUS, University of Santiago de Compostela, Santiago de Compostela, Spain.

Penalised regression splines (P-splines, Eilers and Marx, 1996, *Stat Sci*) models have achieved great popularity both in statistical and in applied research. From a statistical point of view, the reason for their popularity is their applicability to different fields, from survival analysis and spatial and spatio-temporal statistics to functional data analysis. From the applied point of view, the methodological developments have allowed analysing and understanding complex biological and health phenomena. A possible drawback of P-spline models is that they assume a smooth transition of the covariate effect across its whole domain. In some practical applications, however, more complex situations arise, with effects that may not change in some regions of the covariate, while changing rapidly in other regions. In these situations, it is desirable and needed to adapt smoothness locally to the data, and adaptive P-splines have been suggested (e.g., Krivobokova et al., 2008, *J. Comput. Graph. Stat.*). However, the extra flexibility afforded by adaptive P-splines is obtained at the cost of a very high computational burden, especially in a multidimensional setting (e.g., two-dimensional interaction surfaces). Furthermore, to the best of our knowledge, the literature lacks of proposals for adaptive P-splines in more than two dimensions. Motivated by the need of analysing data derived from experiments conducted to study neurons' activity in the visual cortex of behaving monkeys, in this work we present a locally adaptive P-spline model in three dimensions (space and time). Estimation is based on the SOP (*Separation of Overlapping Precision Matrices*) method (Rodríguez-Álvarez et al., 2018, *Stat Comp*), which provides the stability and speed we look for.

Keywords: Penalised splines, adaptive smoothness, visual receptive field.

Integrated models to correct the partial monitoring problem in Ecology

Blanca Sarzo[1], *Ruth King*[2], *David Conesa*[1], *Jonas Hentati-Sundberg*[3]

[1]Blanca.Sarzo@uv.es, David.V.Conesa@uv.es, Department of Statistics and O.R., University of Valencia

[2] Ruth.King@ed.ac.uk, School of Mathematics, University of Edinburgh

[3] jonas.sundberg@slu.se, Department of Aquatic Resources, Swedish University of Agricultural Sciencies

In Ecology, due to the monitoring effort and the inaccessibility of breeding areas, it is often the case that a fraction of the population is monitored. These monitored areas are assumed to be a random sample of the whole population. However, this assumption is often violated. As a consequence, parameter estimates obtained from the monitored areas may be biased (Sanz *et al.*, 2016). This problem is called partial monitoring.

We consider a particular case related to a large capture-recapture-recovery database obtained from the largest colony of Common guillemot (*Uria aalge*) of the Baltic Sea. We model annual survival and resighting probabilities in relation to the age of the individuals through integrated models (Besbeas *et al.*, 2002). In particular, we establish four age categories: 1 (one year old individuals), 2 (two years old individuals), 3 (three years old individuals) and 4 (individuals from four to ten years old).

In this colony, the partial monitoring is present at breeding ledges, where mainly 3 and 4-10 individuals are resighted. As a result, survival probabilities for those age classes are underestimated while recapture probabilities are overestimated. In this study we compare the results obtained from the standard Cormack-Jolly-Seber model and the suitably parameterised integrated model that incorporates recoveries from all individuals in the colony (and not only those at the monitored sites). The results obtained show how the use of the integrated models solve this partial monitoring problem in this colony, although it can be easily extended to other populations.

Keywords: Hidden Markov models, capture-recapture-recovery models.

AMS: 62

Bibliography

Besbeas, P., Freeman, S. N., Morgan, B. J. T. and Catchpole, E. A.(2002). Integrating Mark-Recapture-Recovery and Census Data to Estimate Animal Abundance and Demographic Parameters. *Biometrics 58: 540-547.*

Sanz-Aguilar, A., J.M. Igual, D. Oro, M. Genovart and Tavecchia, G.(2016). Estimating recruitment and survival in partially monitored populations. *Journal of Applied Ecology 53(1): 73-82.*

Generalized partially linear models on Riemannian manifolds

A. Simó[1], *M.V. Ibáñez*[1], *I. Epifanio*[1],*V. Gimeno*[1]

[1]Department of Mathematics-IMAC, Universitat Jaume I. 12071-Castelló. Spain.
Corresponding e-mail:mibanez@uji.es

The problem of predicting a categorical or ordinal variable as a function of a set of covariates (supervised classification) arises in many real-life situations, and it has been widely studied when the covariates lie on a Euclidean space. But when predictive variables, such as the shape of an object, lie on a Riemannian manifold, supervised learning techniques are not so well suited.

This work is motivated by an experimental study carried out by the Biomechanics Institute of Valencia (IBV), whose ultimate objective is to implement a web application for online shopping for children's wear. In particular, that application should make it possible to find the garment's size that best fits a given child, without requiring the child to try on the clothes.

To address this problem, the IBV scanned a sample of Spanish children, getting each body shape represented by 3075 3-dimensional landmarks. Then, a subsample of these children performed an additional fit test, trying on the same shirt in different sizes: the supposedly correct size, the size above and the size below. Then, an expert in clothing and design qualitatively evaluated the fit in each case as too small, correct fit or too large. The aim of the work is to show how generalized partially linear models can be used to predict the goodness of fit of a given garment size, as too small ($\mathrm{Y}_i = 1$), correct fit ($\mathrm{Y}_i = 2$) or too large ($\mathrm{Y}_i = 3$), as a function of the garment size, the (centroid) size of the child and his/her shape (defined on the Kendall's Shape Space). Partially linear models are regression models in which the response depends on some covariates linearly but on other covariates nonparametrically. They generalize standard linear regression techniques and are special cases of additive models, which makes it easier to interpret the effect of each variable. Our aim is to define a generalized partially linear model (for an ordered ordinal response) on Riemannian manifolds (in particular on the Kendall's Shape Space), to develop and illustrate the algorithms for estimating it and to apply it to the children's garment fit problem.

Keywords: Kendall's shape space, partially linear models, generalized linear models.

Assessing the predictive performance of the Cariogram

Mario Trottini[1], *Guglielmo Campus*[2], *Denise Corridore*[3], *Fabio Cocco*[4],*Maria Grazia Cagetti*[5], *Isabel Vigo*[6], *Polimeni Antonella*[7], *Maurizio Bossú*[8]

[1]mario.trottini@ua.es, Department of Mathematics, University of Alicante

[2]gcampus@uniss.it, Zahnmedizinische Kliniken (ZMK) University of Bern and Department of Surgery, Microsurgery and Medicine Sciences, School of Dentistry, University of Sassari

[3]denise.corridore@uniroma1.it, Department of Oral and Maxillofacial Sciences, 'Sapienza' University of Rome

[4]fcocco@uniss.it, Department of Surgery, Microsurgery and Medicine Sciences, School of Dentistry, University of Sassari

[5]maria.cagetti@unimi.it, Department of Biomedical, Surgical and Dental Science, University of Milan

[6]vigo@ua.es, Department of Applied Mathematics, University of Alicante

[7]Antonella.Polimeni@uniroma1.it, Department of Oral and Maxillofacial Sciences, 'Sapienza' University of Rome

[8]maurizio.bossu@uniroma1.it, Department of Oral and Maxillofacial Sciences, 'Sapienza' University of Rome

Several epidemiological studies in the last decade revealed that an increasing high percentage of caries lesions is confined to a minority of high risk children. Such disease pattern has motivated the development of caries risk assessment programmes to identify individuals with higher caries risk in order to involve them in preventive program measures. Among these programmes one of the most popular is the Cariogram. This is a computer program, based on known literature, that models the risk of new caries as a function of nine caries-related factors and their interaction. Caries risk is calculated through an algorithm and expressed as probability of avoiding new caries. These probabilities are used to classify subjects into five risk groups, from low to very high and decide whether or not preventive treatment is necessary. Despite its popularity, the study of the predictive ability of the Cariogram is only partially understood. The aim of this work is to highlight some important limitations of current assessment and reporting of the Cariogram predictive performance and suggest tools and guidelines to overcome these limitations. The predictive performance of the Cariogram is investigated by comparing the Cariogram predictions and observed new caries developments in a two-year follow-up study with Sardinian school children. The results are compared with relevant scientific literature on caries research. Although our discussion in this work is limited to the Cariogram, the ideas discussed are relevant for the predictive performance of other caries risk assessment models that returns probabilistic forecasts for a binary event, such as, for example the National University of Singapore Caries Risk Assessment (NUS-CRA) model or the recent version of the Caries Management by Risk Assessment (CAMBRA) model.

Keywords: Caries risk assessment, Cariogram.

AMS: 62M20

Constructing a deprivation index at a census tract level from the Spanish national census of 2011

Carlos Vergara-Hernández[1],*Miguel Ángel Martínez-Beneito*[2]

[1]vergara_car@gva.es, Fundación FISABIO

[2]martinez_mig@gva.es, Fundación FISABIO

Socio-economic deprivation is a well known health determinant intended to capture the lack of social and/or economical resources. Although the impact of deprivation has been exhaustively studied it has an important drawback since it cannot be directly measured for any population. As a consequence deprivation is usually indirectly quantified from sets of related variables that would be supposed to depend on, or to show the effect of, deprivation. It is therefore not surprising that many different information sources and statistical tools have been used in order to calculate deprivation indexes.

National population censuses have been frequently used for building deprivation indexes at small area levels, such as census tracts. In this case, multivariate statistical techniques are used for extracting the common underlying factor to several variables that would depend on deprivation. The MEDEA, a collaborative project for studying mortality in large Spanish cities, used a simple Principal Component Analysis some years ago for building a deprivation index for census tracts.

The aim of this work is to continue the work started in MEDEA and to generate a new proposal for the construction of a deprivation index in large Spanish cities at a census tract level, but for the 2011 Spanish national census. The main difficulty arising for the use of this new census is that it is based on a sample instead of having all the population available, such as for the 2001 census. As a consequence the new census has additional noise that should be ideally taken into account for this new depriviation index. However, the sampling fraction is unknown for each census tract, what makes our goal even more complex.

In this work we introduce our methodological proposal for building a new MEDEA Project deprivation index for 2011, by considering (i) spatial dependence between nearby locations (ii) temporal dependence between the 2001 and 2011 indexes and (iii) solving the sampling fraction issue previously mentioned.

Keywords: Deprivation index, spatial statistics, multivariate statistics, census data.

AMS: 62H11, 62H25

A comparison of variable selection methods in high-dimensional survival analysis: an application to professional sports injuries

Lore Zumeta Olaskoaga[1], *Jon Larruskain*[2], *Josean Lekue*[2], *Eder Bikandi*[2], *Igor Setuain*[3,4] *and Dae-Jin Lee*[1].

[1]lzumeta@bcamath.org, dlee@bcamath.org,
BCAM-Basque Center for Applied Mathematics,
[2]j.larruskain@athletic-club.eus, j.lekue@athletic-club.eus, e.bikandi@athletic-club.eus,
Medical Services, Athletic Club,
[3,4]isetuain@tdnclinica.es,
Public University of Navarra & TDN-Orthopaedic Surgery and Advanced Rehabilitation.

In sports medicine, injury prediction is one of the most challenging issues. Among other tests, clinicians use multifactorial screening tests in order to gain knowledge in the player's health status and his/her injury risk. In each screening, a large number of measurements are collected. As a result, the data collected have the characteristic that the number of covariates dramatically exceeds the number of observations –also known as high-dimensional data. Consequently, the statistical standard approaches to examine the complex relationships between multifactorial screening tests and injury occurrence cannot be applied directly.

We present some methods for survival analysis in high-dimensional settings, and we show the main pros and drawbacks of each approach. We apply each of the introduced methods to a professional football team with 24 female players, which during the 2017-2018 season completed two multifactorial prevention screening tests. A total of 21 injuries were recorded during the season. We finish with a comparative study where we determine a group of covariates that mainly explain the risk of injury.

Keywords: High-dimensional data, survival analysis, sports injury prediction.

Contribuciones Poster

Comparing penalized ML estimation approaches in a case-control study with a small sample size and more than one covariate: illustration based on simulations and Catalan infants' pertussis data

Lesly Acosta[1,2], *Gloria Carmona*[2], *Carmen Muñoz*[3,4], *Mireia Jané*[2,4]

[1]lesly.acosta@upc.edu, Universitat Politècnica de Catalunya/BarcelonaTech

[2]Public Health Agency of Catalonia, Barcelona

[3]Institut de Recerca Pediàtrica, Hospital Sant Joan de Déu, Barcelona

[4]CIBER Epidemiology and Public Health

With small samples, logistic regression estimates obtained via ML estimation may be biased. If, additionally, the so called separation problem arises, unreliable estimates are obtained (Mansournia *et al.*, 2017) and alternative procedures are needed. In a previous work, various approaches to estimate the odds ratio (OR), including the Firth and the Log-F(m,m) methods (Rahman & Sultana, 2017), were compared with the classic ML estimation. Results showed that the Firth-type approach outperformed the counterpart approaches when using only one covariate (Acosta *et al.*, 2018). It is not clear however, if the same results are attained in settings with more than one covariate.

The motivation for this work is based on a real-data situation under PERTINENT (Pertussis in Infants European Network) set up by the European Center for Disease Prevention and Control, and in which 37 hospitals and 7 study sites participate. The aim of this network is to estimate the disease burden of pertussis as well as the vaccine effectiveness ($\text{VE} = (1 - \text{OR}) \cdot 100\%$, where OR is the odds ratio associated to vaccine) in infants younger than one year.

In Catalonia, PERTINENT was launched in January 2016 and is coordinated by the Catalan Public Health Agency including patients from the Sant Joan de Deu Hospital (HSJD). Cases are eligible infants younger than one year attended at the HSJD, testing positive for Bordetella pertussis. Using a test negative design, for each case, 3 controls are randomly recruited.

The objective of this work is to compare different penalized ML approaches with small samples and taking into account more than one covariate. For this purpose, an exhaustive simulation study is carried out and the different methods are illustrated with the Pertussis data collected in Catalonia. All methods are implemented in R.

Keywords: Case-control study, logistic regression, penalized likelihood estimation.

Variable selection in functional regression: application in graft-versus-host disease

M. Carmen Aguilera-Morillo[1], *Ismael Buño*[2], *Rosa E. Lillo*[3], *Juan Romo*[4]

[1]maguiler@est-econ.uc3m.es, Department of Statistics, University Carlos III de Madrid

[2]ismaelbuno@iisgm.es, Department of Hematology, Genomics Unit, Gregorio Marañón General University Hospital and Gregorio Marañón Health Research Institute (IiSGM)

[3]lillo@est-econ.uc3m.es, UC3M-BS Institute of Financial Big Data, Department of Statistics, University Carlos III de Madrid

[4]romo@est-econ.uc3m.es, Department of Statistics, University Carlos III de Madrid

LASSO is one of the most extended techniques for variable selection. In that sense, an approach for functional LASSO in terms of basis representation of the sample paths, related to the response variable, is proposed. This problem is motivated by a real application in the graft-versus-host disease, which is the main complication (30-50%) after allogeneic hematopoietic stem-cell transplantation and the most important cause of non-relapse mortality.

Allogeneic hematopoietic stem-cell transplantation (allo-SCT) is a curative therapeutic approach for patients with hematologic malignancies. Patients undergoing allo-SCT receive a donor graft containing hematopoietic stem cells, as well as various other cell types, including alloreactive T cells. T cells promote hematopoietic engraftment, T-cell immunity reconstitution and mediate graft-versus-leukemia effect, which may prevent tumor relapse. However, donor T cells may also cause graft-versus-host disease (GVHD).

The aim of this study is the selection of the clinical variables that are most related to the genotype of patients affected by chronic GVHD. The information on the patient genotype is provided in terms of SNP (single-nucleotide polymorphism) data, which has been transformed into a functional data set with the appropriate techniques.

This dataset is part of a wider study which was approved by the ethics committee of Hospital General Universitario Gregorio Marañón, and all recipients and donors provided written informed consent according to the Declaration of Helsinki.

Keywords: Graft-versus-host disease, function-on-scalar regression, LASSO.

Polygenic risk scores for child-onset psychiatric disorders and cognitive trajectories in schoolchildren

Sofia Aguilar-Lacasaña[1], *Natalia Vilor-Tejedor*[2], *Jordi Sunyer*[3], *Silvia Alemany*[4]

[1]sofia.aguilar@isglobal.org, Barcelona Research Institute for Global Health (ISGlobal). University of Vic - Central University of Catalonia (UVIC-UCC).

[2]natalia.vilortejedor@crg.eu, Centre for Genomic Regulation (CRG). The Barcelona Institute for Science and Technology. Barcelonabeta Brain Research Center (BBRC), Pasqual Maragall Foundation.

[3]jordi.sunyer@isglobal.org, Barcelona Research Institute for Global Health (ISGlobal). Universitat Pompeu Fabra (UPF). CIBER Epidemiología y Salud Pública (CIBERESP). IMIM (Hospital del Mar Medical Research Institute).

[4]silvia.alemany@isglobal.org, Barcelona Research Institute for Global Health (ISGlobal). Universitat Pompeu Fabra (UPF). CIBER Epidemiología y Salud Pública (CIBERESP).

Results from the most recent genome-wide meta-analysis on attention-deficit/hyperactivity disorder (ADHD) suggest that multiple genetic variants of small effect contribute to the aetiology of the disorder, implying a highly polygenic architecture (Demontis et al. 2018). The core symptoms of ADHD, inattention and hyperactivity, are closely related to cognitive difficulties such as inattentiveness, which are common among ADHD patients (Frazier et al. 2004). However, the extent to which these difficulties may be explained by genetic risk for ADHD remains largely unknown. Herein, we use a polygenic risk score (PRS) approach (PRSice) to investigate the genetic overlap between ADHD and inattentiveness trajectories in schoolchildren. Data on genetics, covariates and inattentiveness were obtained from 1643 children (6185 observations) drawn from the BREATHE project (Sunyer et al. 2015). We used linear mixed-effects models including children nested within schools as random effects to account for the multilevel nature of the data. To capture the change in the inattentiveness trajectories associated to PRS, an interaction term with age and PRS was included. Models were adjusted by age, sex, maternal education and socioeconomic status. No significant associations between ADHD-PRS and inattentiveness trajectories were found. These results indicate that genetic variants related with ADHD do not account for variation in attention performance among schoolchildren. Our results are in line with previous research reporting no consistent evidence of shared genetic effects between ADHD and cognitive function (Clarke et al. 2015). A potential explanation may be that worse attention performance is a consequence of the symptoms and not genetically drive. Next steps include examining the genetic overlap between autism spectrum disorder (ASD) and cognitive trajectories. Both ADHD and ASD are child-onset psychiatric disorders, highly heritable and polygenic and individuals affected show cognitive deficits. Thus, it is interesting to elucidate whether cognitive deficits in these disorders have different etiological origins.

Keywords: ADHD, cognitive trajectories, polygenic risk score, inattentiveness.

Bayesian radiocarbon dating for understanding demographic cycles in the Iberian Neolithic

Carmen Armero[1], *Gonzalo García-Donato*[2], *Salvador Pardo*[3], *Joan Bernabeu*[4]

[1]carmen.armero@uv.es, Departament d'Estadística i Investigació Operativa, Universitat de València

[2]gonzalo.garciadonato@uclm.es, Departmento de Economía y Finanzas, Universidad de Castilla-La Mancha

[3]salvador.pardo@uv.es, Departament de Prehistória, Universitat Autónoma de Barcelona

[4]juan.bernabeu@uv.es, Department de Prehistória, Arqueologia i Història Antiga, Universitat de València

Bayesian inference for chronological models focuses on the analysis and interpretation of chronological information from scientific data as well as expert knowledge. We construct a chronological model for the Iberian Neolithic (from c. 5600-4000 cal. BC) based on the main archeological sites in Eastern Mediterranean Spain with a good chronological stratigraphy - Cova de les Cendres and Cova de l'Or- that provide a time axis of consecutive phases of the general period of the study. Together with these two sites, we use information from other 11 sites with short stratigraphic development. Relationship between these sites and the Cendres-Or stratigraphic time-axis is made on similarities of ceramic decorations.

The Bayesian model starts with an informative prior distribution for the true calendar dates of the different phases involved which accounts for the stratigraphic and ceramic knowledge. Data include simple radiocarbon determinations for the different sites and phases and associated laboratory errors. The joint posterior distribution for the subsequent true calendar ages is approximated by means of Markov chain Monte Carlo methods methods through WinBUGS software.

Keywords: Chronological model, radiocarbon determination, stratigraphic information.

Multivariate analysis as a tool in the optimization of soil quality indexes for the semiarid ecosystem of the Northern Plateau, Spain

C. Ávila-Zarza[1], *F. Santos-Francés*[2], *A. Martínez-Graña*[3], *M. Criado*[2], *Y. Sánchez*[3]

[1]caaz@usal.es, Department of Statistics, University of Salamanca.

[2]fsantos@usal.es, Department of Soil Sciences, University of Salamanca.

[3]amgranna@usal.es, Department of Geology, University of Salamanca

World population growth and agricultural expansion are among the main causes of soil degradation in most terrestrial ecosystems. Therefore, soil quality evaluation has gained widespread interest lately, as a way to protect and preserve soil. The improvement of the evaluation of the soil quality is imperative for the development of sustainable agriculture and can also be used to evaluate the sustainability of land. Quality indexes are obtained by integrating different soil property indicators, which provide information on soil. However, their suitability and implementation in large areas is difficult because it requires to obtain a large number of soil variables. Therefore, it is important to develop quality evaluation methods that use a minimum number of indicators to improve their work efficiency and reduce labour time and expenses of research; in this regard, we use multivariate dimensionality reduction techniques to achieve it The present study was carried on in the semiarid ecosystem of the Northern Plateau (Spain) that covers a total area of 770 km2. To characterise this area, 300 samples were collected from 75 soil profiles considering the soil surface properties (between 0 and 25 cm depth) in addition to the properties of all horizons of the soil profile (between 0 and 100 cm), where we evaluate the quality of different soil types, with different land use types. Using a statistical multivariate approach, fundamentally through the use of Factorial Analysis, we establish the most suitable quality index models for this region. This procedure is particularly relevant given that the use of a limited number of soil indicators reduces the analysis cost and therefore, allows increasing the sampling density for large-scale evaluations.

Keywords: Factorial analysis, soil quality index.

AMS: 62H25, 62P12

Archetypoid analysis applied to the teaching of mathematics

Ismael Cabero[1], *Irene Epifanio*[2]

[1]ismael.cabero@uv.es, Dept. de Didàctica de la Matemàtica, Universitat de València
[2]epifanio@uji.es, Dept. Matemàtiques-IMAC, Universitat Jaume I

One of the first and most crucial problems to solve in universities is the newcomers' lack of preparation. In this work we have binary answers to a mathematical questionnaire that was made to first year students of the Jaume I University of Castelló (UJI). We want to look for the best way to classify these students in order to support those who have similar profiles and solve the problems they share. A good screening and selection in different groups is central. The most appropiate method for finding relationships and structures in order to facilitate the understanding and interpretation of the different groups will be the archetypal analysis (ADA).

Archetypal analysis is an exploratory tool that explains a set of observations as mixtures of pure (extreme) patterns. If these patterns are convex combinations of observations, we refer to them as archetypes. But, if the patterns are actual observations of the sample, we refer to them as archetypoids. ADA uses the virtues of the archetypes, adding more intelligible results using representatives of the same sample. We will demonstrate the appropriateness of using these real representatives and compare the result with other more established statistical techniques such as homogeneity analysis (HOMALS), cluster analysis (PAM) and probabilistic archetype analysis (PAA). We will also change the frame of reference and find the archetypoids of the questions in order to find relationships between them.

This perspective is absolutely innovative within the ADA and it will also be the first time that ADA is used for a set of binary data, which opens a wide range of possibilities to other studies.

Acknowledgments: This work is supported by grants DPI2017-87333-R from the Spanish Ministry of Science, Innovation and Universities and UJI-B2017-13 from Universitat Jaume I.

Keywords: Arquetypal analysis, binary observations, didactic of mathematics.

Analysis of human microbiome data in the process of ageing using GAMLASS models

Enrique Calderin[1],*Guillermo H. Lopez-Campos*[2],

[1]enrique.calderin@unimelb.edu.au, Centre for Actuarial Studies, The University of Melbourne.
[2]g.lopezcampos@qub.ac.uk, Wellcome-Wolfoson Institute for Experimental Medicine, Queen's University Belfast

The microbiome and its changes under different conditions and time is an important area of biomedical research as it has been linked as a relevant factor associated with the risk of suffering an increasing number of health conditions. The microbiome is a dynamic component that undergoes different changes depending on many factors such as diet, medications or age. In the case of the latter, it has been observed that changes in composition of the microbiome are more correlated with the biological age than with the chronological age. Therefore, the analysis of the microbiome may lead to the identification of relevant biomarkers that can be associated with life expectancy. In this work, using an already published dataset we aimed to analyse the application of different statistical methods to identify the relationships between the changes in the microbiome during the ageing process and other metadata. For this purpose, we have used a cross-sectional dataset built up with 16S RNA sequencing data and metadata from more than 300 Japanese volunteers of different ages. We focused in the application of GAMLASS (Generalized Additive Models for Location, Scale and Shape) models as this approach allowed us to apply different data transformations of the location, scale and shape parameters. The flexibility of this methodology allowed us to explore the use of both discrete and continuous probability distributions for the response variable. In addition, it allowed us to use a large amount of additive penalty elements in our analyses. Finally, the algorithms used are modular and allow us to easily include new elements in the generation of the predictor that might be associated with different sources or the different phyla associated with the microbiome analyses. We have used statistical methods and approaches to analyse microbiome evolution associated with ageing process. Some of the aims of this project are related with the possibility of applying these methods in combination with other relevant data sources (such as life expectancy) to assess the possibility of applying microbiome information as a surrogate for biological age and use it for risk assessment for insurance policies

Keywords: Microbiome, GAMLASS, statistics.

AMS: 62P10, 62P05

Factors associated with family resilience under food restriction in Central America and Mexico, estimated by multinomial logistic regression

Camacho-Sandoval, Jorge[1], *Romero-Zúñiga, Juan José*[2]

[1]jorge.camacho.s@gmail.com, Epidemiology Postgraduate School, Universidad Nacional de Costa Rica.
[2]juan.romero.zuniga@una.cr, Epidemiology Postgraduate School, Universidad Nacional de Costa Rica.

Mesoamerica is a region with large areas of vulnerable populations. To know the strategies of family resilience, a cross-sectional study was conducted through a structured survey, to more than 23 thousand families receiving sponsorship for their children, from a multinational NGO, from Mexico to Costa Rica. The families have an average of 4.2 members (SD 2.3, max 23), 15% with a woman as head, with the highest percentage in Costa Rica (28.5%) and the lowest in Guatemala (0.3%). Three resilience strategies to food restriction in the year prior to the survey were defined: 1) adaptation (ADAP): changes in diet (quantity, quality, frequency); 2) sale of liquid assets (VAL): consumables or non-essential items for productive work; 3) sale of production assets (VAP): tools, vehicles or seeds. In the last two cases, it is understood that this strategy was used to aid the purchase of food. Of the ADAP mechanisms, the most frequent were: eating wild products (40.4%), cheaper diet (31.2%), reducing portions (26.1%), doing extra work (29.5%) and a worrying 12% of school dropouts. In the cases of VAL, borrowing money to eat (21.3%) and eating seeds intended for planting (11.3%), were the most used strategies. Finally, with respect to VAP, asking for credits for planting was the most used strategy, but at only 6.4%. The sale of tools or equipment was not reported, significantly. For the analysis of factors associated with family resilience, the response strategies were grouped into four categories: 0 = no response, 1 = ADAP, 2 = ADAP + VAL or ADAP + VAP, 3 = VAL or VAP. A multinomial logistic model was built, forward stepwise, using the likelihood-ratio as test for the fit of the model. There was a consistent effect of country (p <0.001), of the number of household members sponsored by the NGO (p = 0.007), the level of schooling of the household head (p = 0.007), the number of household members among 5 and 18 years-old (p <0.001), as well as over 18 (p <0.001), and the presence of the spouse (p = 0.004), when comparing strategies 1, 2 and 3 vs. 0. For the age of the head of the family, it was association only when contrast strategy 3 vs. 0 (p <0.034). With Costa Rica as reference, households in Guatemala and Honduras presented less than half the risk of adopting some resilience mechanism, while Mexico and Honduras had between 111% and 46% more risk. There was a clear tendency towards resilience as the level of education of the head of the family declined. Finally, the increase in the number of members between 5 and 18 years-old, as well as over 18, presented increases between 4% and 10% for each increment in one unit. Factors such as the sex of the head of the family or the masculinity index did not show a significant association. The study of factors that lead vulnerable families to take resilience measures, is essential to design the most appropriate support and social assistance strategies by NGO and state entities.

Keywords: Social epidemiology, risk estimation, multinomial logistic regression.

Integrating environmental health characteristics for health techonology assesstment. A systemic approach. The case of the ecosystem in Drake bay, Puntarenas, Costa Rica.

Milena Castro[1][2][3]

[1]milena.castromora@ucr.ac.cr, Escuela de Estadística, Universidad de Costa Rica

[2] Centro de Investigación en Matemática Pura y Aplicada, Universidad de Costa Rica

[3] Centro de Investigaciones en Ciencias Atómicas, Nucleares y Moleculares, Universidad de Costa Rica

Policy making for environmental health implies consideration of a variety of indicators proposed by the World Health Organization and started by Centro de Investigación sobre el Síndrome del Aceite Tóxico y Enfermedades Raras (CISATER). Different observational perspectives can be identified with air, radiation, water, soil, residuals, sanitation, noise, traffic accidents, food safety, infraestructure, ocupational conditions, chemical emergencies and polluted areas. These spatial characteristics can be contrasted with longitudinal community observations of the socio-economic dynamics. However, challenges arise when data available is heterogeneous as comes from a collection of sources of information. A complex model can be defined when integrating more than one conceptual dimension. Dimensions can be specified according to observational techniques: survey data for a socioeconomical and epidemiological characterization of the population, environmental data based on analytic screening of water sources, and clinical epidemiology data can be obtained, in order to elaborate a systemic approach using Markov model simulation. The response of the model is related to the quality of the environment, to identify community development strategies according to its potential and needs satisfaction, like food safety. Model specification allows evaluation of technologies to be implemented at a populational level. Bio-Sand filters were designed and an experimental observation was undertaken with a family in Drake. A decrease in *Escherichia coli* was observed, but termotolerant coliforms had an increase, after comparing before and after bio-sand filtered water samples from Drake's main basin. Evidencing a health policy for Drake's ecosystem implies overtaking microbiological assessments. How an aqueduct should be developed for a population living around areas under forestal and water conservation? Nowadays, this is a relevant research question for Drake's ecosystem, where biodiversity and water resources represent an important component of its turism based economy.

Keywords: Markov models, complex models, water technologies.

A Comparative study of the runs test for the hypothesis of symmetry

Corzo, J.A.[1], *Vergara, M.* [2], *Babativa, G.*[3]

[1]jacorzos@unal.edu.co, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá D.C.
[2]mvergara@unisalle.edu.co, Departamento de Ciencias Básicas, Universidad de La Salle, Bogotá D.C.
[3]jgbabativam@usal.es, Departamento de de Estadística, Universidad de Salamanca, Salamanca .

We present a summary of the main results obtained in (Corzo et al. 2019), about the proposed a trimmed runs test for the hypothesis of symmetry with one sided alternative, in samples coming from the Generalized Lambda Distribution (GLD) with unknown median. The size of the proposed test is calibrated with four symmetrical cases of the GLD and the empirical power is compared with that of some other tests for the same hypothesis, using eight asymmetrical cases of the GLD. Results show that the proposed test is unbiased in the cases used for calibration, and that the empirical power of the proposed test overtakes the empirical power of all compared tests, excepting one of them in two specific cases. Some hints are given on how to optimize the empirical power according to the size of the tails of the sampled distributions. There are many tests for the hypothesis of symmetry with the two-sided alternative. However, for the one-sided alternative there are too few parametric or non-parametric tests. (Babativa & Corzo 2010) use the idea of trimming by (Modarres & Gastwirth 1996) to build a trimmed runs test for the two-sided alternative; later, (Corzo & Babativa 2013) proposed a trimmed J_6-test for the two-sided alternative, weighting the values of the test statistic positively or negatively according to the tail where the observations of the sampled distribution are placed. We will use the positive and negative values of the test statistic to study the behavior of the empirical power of the same test for the one-sided alternatives K_1 and K_2. We will work the test for alternative K_1; the alternative K_2 requires minor modifications. In this context, (Corzo et al. 2019), propose one test and give some distributional properties of the test statistic. In this work we shows the results of a simulation study to compare our test with some other tests.

Keywords: Runs tests, one sided symmetry tests, power of a test.

References

Babativa, G. & Corzo, J. A. (2010), Propuesta de una prueba de rachas recortada para hipótesis de simetría, *Revista Colombiana de Estadística*, **33**(251), 271.

Corzo, J. & Babativa, G. (2013), A modified runs test for symmetry, *Journal of the Statistical Computation and Simulation*, **83**(5), 984-991.

Corzo, J., Vergara, M. & Babativa, G. (2019). A runs test for the hypothesis of symmetry with one sided alternative, *Universitas Scientiarum*, **24**(2).

Modarres, R.& Gastwirth J. L.. (1996), A modified runs test for symmetry, *Statistics & probability letters*, **31**(2), 107-112.

Application of meta-analysis in nitrogen fertilizer studies in Puerto Rico

Alejandra M. De Jesús Soto[1], *Raúl E. Macchiavelli*[2]

[1]alejandra.dejesus@upr.edu, Department of Mathematical Sciences, University of Puerto Rico Mayagüez

[2]raul.macchiavelli@upr.edu, Department of Agroenvironmental Sciences, University of Puerto Rico Mayagüez.

Taking a final decision about the validity of a hypothesis or an estimate should not be based on the results of a single experiment, mainly because results frequently vary across studies. Instead, efforts have been made to develop statistical procedures to synthesize data from one study to another. Meta-analysis combines the evidence from a collection of available studies on a topic of interest or specific question. This tool is built on the principle that individual studies, surveys, and observations contribute to the overall total knowledge base. Most applications of meta-analysis are on to validate the significance of treatment effects in comparative experiments. The main goal of this work is to apply these methods to integrate estimates of nitrogen fertilizer recommendations for Solanaceae crops in Puerto Rico (mainly green pepper, eggplant, and tomato). This is done by combining estimates of nitrogen CNR (Crop Nutrient Requirement) obtained in different experiments under comparable conditions. The CNR for each individual study was obtained by fitting a nonlinear model to the relative yield vs. fertilizer dose relationship. The advantages and disadvantages of mixed model approaches to meta-analysis (ML and REML) are discussed in the context of this set of studies.

Keywords: Mixed models, crop nutrient requirement, CNR

Non-linear mixed models implementation in InfoStat and interface to the nlme and lme4 libraries in R

Julio A. Di Rienzo[1], *Raúl Macchiavelli*[2], *Fernando Casanoves*[3], *Mónica Balzarini*[1]

[1]Facultad de Ciencias Agropecuarias-Universidad Nacional de Córdoba.

[2]Dept. de Cultivos y Cs. Agroambientales, Univ. de Puerto Rico-Mayagüez.

[3]Centro Agronómico Tropical de Investigación y Enseñanza, CATIE.

The *nlme* library of R implements linear and non-linear mixed models through the *lme and nlme* functions. The *lme4* library implements linear, non-linear, and generalized linear mixed models using the *lmer, nlmer,* and *glmer* functions, respectively. For non-linear mixed effects models, the *lme4* library uses Laplace and Gauss Hermite approximations to the likelihood, while the *nlme* library uses a pseudo-likelihood approach. Here we show the implementation, in the framework of InfoStat, of an interface to these functions to fit non-linear mixed effects models. When no random effects are present, the interface uses the *nls* function in a way that is transparent for the end user. Thus, this easy-to-use interface offers a complete non-linear modelling tool to InfoStat users. The implementation is complemented with a tutorial including several worked real life examples. The tutorial presents a short introduction to pseudo likelihood and likelihood estimation for non-linear mixed models, and discusses advantages and pitfalls of these methods using relevant examples. It also includes step-by-step instructions to use several options to obtain adjusted and predicted values as well as graphical tools for diagnostic purposes.

Keywords: Random effects, likelihood estimation, pseudo likelihood estimation.

Classification methods to determine competent resolvers through neuroimaging

Lara Ferrando[1], *Noelia Ventura-Campos*[2], *Irene Epifanio*[3]

[1]lferrand@uji.es, Grup Neuropsicologia i Neuroimatge Funcional, Universitat Jaume I
[2]venturan@uji.es, Dept. Educació i Didàctiques Específiques, Universitat Jaume I
[3]epifanio@uji.es, Dept. Matemàtiques-IMAC, Universitat Jaume I

Problem solving is one of the core elements of math learning. An important line of research in the resolution of verbal problems is the study of the cognitive processes when subjects translate problems into the language of algebra. Typically, when students recognize the information in the statement, but are not able to build a correct equation, this would be known as a reversal error (RE).

The experimental work performed, shows a neuroimaging study in which data are collected through magnetic resonance images (MRI) while performing a verbal problem-solving task with RE with a total of 20 participants. To obtain the RE and non-RE groups, we used the answers obtained during the MRI. The statistical analyses of MRI were performed by SPM12 using the General Linear Model for each participant and for each time point. Finally, the classification of the groups as RE or non-RE, was carried out with the R program, where the activated brain areas were used as variables. We tested 13 classification methods, using leave-one-out cross-validation.

After testing with 13 classifiers, it can be concluded that the methods that best classify in our case, taking into account that it is a small sample, would be the flexible discriminant analysis corresponding to an 80% success rate, followed by linear discriminant analysis and logistic regression, which have the same percentage of success (70%). These are simple methods, as suggested by Hand, who said that on many occasions the simplest classical methods could work better than more recent and sophisticated methods, due to uncertainties and arbitrariness, and that this could especially occur in real problems. Moreover, the results show that the brain areas activated and introduced as classification variables could be considered good biomarkers to help us identify competent resolvers.

Acknowledgments: This work was supported in part by Universitat Jaume I Grant UJI-A2017-8.

Keywords: Statistical classification, general linear model, neuroeducation.

Mortality risk of prioritized diseases in the Colombian Caribbean region for the 2008-15 period. Mapping and analysis from a Bayesian approach

Karen Cecilia Flórez Lozano[1], *Edgar Navarro Lechuga*[2]

[1]lozanok@uninorte.edu.co, Department of Math and Statistics, Universidad del Norte.
[2]enavarro@uninorte.edu.co, Department of Public Health, Universidad del Norte.

The first causes of diseases in the world are non-communicable diseases. With an ecological study, the mortality data available in the database of the National Administrative Department of Statistics (DANE) were studied. Mortality rates were adjusted using standardized indexes (REM) by age and gender, and using Bayesian methods, the risk of death due to prioritized diseases was estimated. We evaluate the temporal evolution with adjusted annual mortality rates. Between 2008 and 2015 approximately 58% (86,185 deaths) were due to disorders in the circulatory system, followed by malignant tumors with 24.4% (36,188 deaths). The first place is occupied by ischemic heart disease, with a significant increase ($p < 0.0001$) in risk starting in 2011, as well as for prostate cancer (p value < 0.0001) and breast neoplasms in women (value of $p = 0.0221$). During the study period, mortality from chronic diseases increased. A greater tendency of risk for men in these diseases is observed, which generates information so that the decision-makers in health care adjust the programs and services of prevention, promotion, care and rehabilitation of diseases in accordance with the reality of the territorial entity.

Keywords: Risk maps, mortality, Bayesian analysis.

RAAPA Shiny App for Risk Assessment around Pollution Sources

Virgilio Gómez-Rubio[1], *José Luis Gutiérrez-Espinosa*[1], *Francisco Palmí-Perales*[1], *Rebeca Ramis-Prieto*[2,3], *José Miguel Sanz-Anquela*[4], *Pablo Fernández-Navarro*[2,3]

[1]Virgilio.Gomez@uclm.es, Department of Mathematics, Universidad de Castilla-La Mancha.

[2]Environmental and Cancer Epidemiology Unit, Carlos III Institute of Health.

[3]Consortium for Biomedical Research in Epidemiology & Public Health, CIBER Epidemiología y Salud Pública - CIBERESP.

[4]Cancer Registry and Pathology Department, Hospital Universitario Príncipe de Asturias and Department of Medicine and Medical Specialties, Faculty of Medicine, University of Alcalá.

Risk assessment around point sources is an important problem in spatial epidemiology. This is often tackled by case-control studies in which exposure to the pollution source is accounted for. For example, when assessing the impact of polluting industries it is customary to take into consideration the distance to the source as a risk factor in the model. When the actual locations of cases and controls are known, a number of models have been proposed to assess the impact of the pollution source on the spatial distribution of the cases after accounting for the spatial distribution of the controls. However, these types of analyses often require complex data handling, such as dealing with map projections, visualization of spatial data, computation of spatial statistics and others.

For this reason, we have developed a Shiny App called RAAPS (Risk Assessment App around Pollution Sources) to help public health practitioners in the analysis of case-control data for risk assessment. The App has a simple interface to upload data (cases, controls, the locations of the hazardous point sources and, possibly, the boundaries of the study region). Maps with the locations of cases and controls, and the study region, are easily created. Similarly, a statistical analysis to assess how the distance to the pollution source impacts the locations of the cases is done. Hence, this could be a usefull tool for public health surveillance at hospitals and Cancer Registries. The use of this App is illustrated on a dataset with the locations of different types of cancer (lung, stomach and kidney) in Alcalá de Henares (Madrid, Spain) and several polluting industries close to the city are used in the analysis.

Keywords: Point patterns, spatial epidemiology, spatial statistics.

AMS: 60G55, 62H11

Global Sensitivity Analysis of complex models combining Morris screening method and variance decomposition method in a robust way

Dorleta García, Inmaculada Arostegui[1]*, Raúl Prellezo*

[1]inmaculada.arostegui@ehu.eus, Departamento de Matemática Aplicada, Estadística e Investigación Operativa, UPV/EHU. BCAM- Basque Center for Applied Mathematics

Variance decomposition methods are the most accurate methods available to conduct GSA. However, they are computationally intensive and the application to complex simulation models with many input factors is usually unafordable. In such cases, a common approach is to first apply a screening method to identify the less important input factors and then apply the variance decomposition method fixing the less important factors. However, the combination of both methods in multi-dimensional output models is not ususally done in a robust way. In general, the selection of the less important factors in the screening method is not optimal and its convergence is not evaluated.

To obtain a robust combination of both methods we define two new criteria, a selection criterion that simulates the visual selection of the input factors and a convergence criterion that ensures that the factors selected will not change if the number of iterations in the screening method is increased. The performance of the criteria was assessed using a complex fisheries management bio-economic simulation model. The Morris screening method needed 200 paths to converge. Furthermore, the performance of the selection criterion was compared with a criterion that selects a fixed number of factors per indicator and a selection criterion based on savage scores. Our proposal outperformed the fixed number of factors criterion. Moreover, the criterion based on savage scores performed better when the results indicators were considered globally but the proposal worked better when they were considered individually

Keywords: Complex simulation model, global sensitivity analysis, Morris screening method.

Mapping malaria relative risk in Colombia. A Bayesian approach using zero-inflated models and intrinsic CAR prior specification

Jose Enrique Gómez Gómez[1], *Marlon Ricardo Ruiz Fernandez*[2], *Carlos Eduardo Melo Martínez*[3]

[1]jegomezg@correo.udistrital.edu.co, Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas.

[2]mrruizf@correo.udistrital.edu.co, Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas.

[3]cmelo@udistrital.edu.co, Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas

This paper aims to estimate the relative risk of malaria in Colombia based on the registered cases for each administrative unit, namely municipalities, for 1998, year in which this disease reach its historical high in the country. Despite that fact, the data present a very large amount of cero counts (83% of the total counts) due to non-endemic areas, leading to a zero-inflated problem analysis. Further assays also suggest the presence of over-dispersion in the non-zero part of the data. A small set of well-known malaria descriptive environmental covariates is take in account for the models estimation as well as the quantification of its corresponding relative risk factors. To face the zero inflation of the data a Zero Inflated Poisson (ZIP) model is proposed, finding lack of fit evidence due to data skewness, to accommodate this problem Zero Inflated Negative Binomial (ZINB) is implemented. Both models were shown to property deal with le large amount of cero counts, however the over-dispersion of the non-zero data proves to be a significant property that must be consider into the count distribution assumed in order to get the best possible model. The spatial correlation is modelled through implementation of General Additive Models (GAM) with a structured spatial random effect specification based on the Intrinsic Conditional Auto-Regressive (iCAR) model proposed by Besag et al. (1991). The estimation of the model is carried out with empiric Bayesian method of Integrated Nested Laplace Approximations (INLA). Finally, relative risk maps are generated out of each proposed model and compared to maps constructed with classical epidemiology risk measures as the Standardised Incidence Rate (SIR) showing that the ZINB model gives the most consistent risk estimation.

Keywords: Malaria, Disease mapping, CAR models, Bayesian approach.

Assessing adaptation to extreme weather events in the city of Valencia

Carmen Iñiguez[1], *Francisco J. Santonja*[2], *Ana Corberán*[2],
Ferran Ballester[3], *Aurelio Tobías*[4]

[1]carmen.iniguez@uv.es. Department of Statistics and Operational Research. Universitat de València. Spanish Consortium for Research on Epidemiology and Public Health (CIBERESP).

[2]francisco.santonja@uv.es; ana.corberan@uv.es. Department of Statistics and Operational Research. Universitat de València.

[3]ballester_fer@gva.es. Department of Nursery and Chiropody, Universitat de Valéncia.

[4]aurelio.tobias@gmail.com. Spanish Council for Scientific Research (CSIC), Barcelona.

In the context of a warming climate, a deep understanding of temperature related health effects is crucial. Research on this topic is usually based on time series analysis and distributed lag-non lineal models (dlnm) have been consistently applied to large data sets to address key issues on this topic. Dlnm (Gasparrini et al 2011) is a class of models able to describe the complex nonlinear and lagged dependencies typically found for temperature and mortality through a bi-dimensional exposure-lag response surface. One of the issues assessed recently has been the attenuation in human vulnerability to heat, (Gasparrini et al 2016). However, a decrease in cold related mortality is less clear (Arbuthnott, et al 2014).

Our aim was to examine the change over time of temperature related mortality in the city of Valencia from 1990 to 2014. With this aim, we analyzed daily death counts and daily mean temperature from the 52 provincial capital cities in Spain. We applied dlnm on quasi-Poisson regression models adjusted by long-term trend and seasonality in each city for those periods spanning 11 years moving across the study period. Local curves were pooled at the national level through multivariate meta-analysis, allowing so that smaller cities learn from the bigger ones. Blups for Valencia were derived and adaptation was assessed in terms of changes in: Temperature associated to the minimum mortality, Relative Risk at 5% and 95% percentiles, triggering points for heat and cold effects and attributable deaths to heat and cold. This study as a procedure for the systematic update of the evidence to be used in the Prevention plan against heat waves in the Valencian Community was awarded by the Medical Valencian Institute and the Council of Valencia.

Keywords: Time series, dlnm, temperature related mortality.

Bayesian hierarchical spatio-temporal models to identify the European hake (Merluccius merluccius) recruits dynamic in the northern Iberian Peninsula

Francisco Izquierdo[1], *Maria Grazia Pennino*[2], *David Conesa*[3], *Iosu Paradinas*[4], *Santiago Cerviño*[1], *Francisco Velasco*[4], *Alexandre Fernández*[5], *Fran Saborido-Rey*[5]

[1]franizq3@gmail.com, david.v.conesa@uv.es Departament d'Estadística i Investigació Operativa, University of Valencia

[2]grazia.pennino@ieo.es, santiago.cerviño@ieo.es, Instituto Español de Oceanografía (IEO), Vigo

[3]3paradinas.iosu@gmail.com, Ipar Perspective Asociación, Karabiondo Kalea.

[4] fran.velasco@ieo.es, Instituto Español de Oceanografía (IEO), Santander.

[5]alex@iim.csic.es, fran@iim.csic.es, Instituto de Investigaciones Marinas (IIM-CSIC), Vigo

Spatial management of commercial resources is becoming an effective measure to be broadly implemented in the European Seas. However, it is currently unconnected from the population dynamics and the temporal assessment. Indeed, it is known that species abundance can be influenced by the environmental features of its own habitat and/or by biotic process that are spatially structured (e.g. reproduction, predation, among others). Usually, this variability is assumed to be implicitly in the abundance trends used as inputs of the stock assessment models and it is not explicitly taken into account. Within this context, in this study we propose a novel methodological approach for an effective implementation of spatial and ecological knowledge that could help to embrace species spatial management providing a more holistic and ecosystem-based approach. As case study we used the European hake recruits (*Merluccius merluccius*) in the northern continental shelf of the Iberian Peninsula. Data collected during the scientific survey series "DEMERSALES" by the "Instituto Español de Oceanografía" (IEO) from 2005 to 2016 were analyzed using Bayesian hierarchical hurdle spatial-temporal models, considering as environmental variables Sea Surface Temperature, Sea Surface Salinity, bathymetry and rugosity of the seabed. The presented models allowed to solve data issues such as spatial dependence, temporal autocorrelation, high proportions of zeros and big size of samples. In addition, using the integrated nested Laplace approximation (INLA) method different types of spatio-temporal structures (persistent, opportunistic, and progressive) were compared in order to achieve the most robust fit for our data type. We argue that the analytical framework proposed in this study allowed (1) to assess which environmental factors influence hake recruits abundance in the northern continental shelf of the Iberian Peninsula, (2) to identify areas in which individuals of this functional group are more aggregated as well as their space-time fluctuations, and (3) to provide improved habitat-based standarization abundance indexes for stock assessment models.

Keywords: Spatio-temporal models, marine ecology, stock-assessment.

A novel two-step approach for the joint modelling of longitudinal and survival outcomes

Valeria Leiva[1], *Susana Eyheramendy*[2], *and Danilo Álvares*[3]

[1]vjleiva@mat.uc.cl, Department of Statistics, Pontificia Universidad Católica de Chile
[2]susana@mat.uc.cl, Department of Statistics, Pontificia Universidad Católica de Chile
[3]dalvares@mat.uc.cl, Department of Statistics, Pontificia Universidad Católica de Chile

In survival biomedical studies it is usual to collect measurements of different biomarkers of the same individual over time. These repeated measurements can be dependent of a survival process. In this case, a popular way of modelling the longitudinal and the survival data is through a joint model approach. Typically, this joint model will consist of two submodels: one for the longitudinal data and the other one for the survival data. The submodel for the longitudinal data is usually a mixed model with random and fixed effects. For the survival data model, there are several approaches that can be considered, e.g. Cox proportional hazard and accelerated failure time (AFT). In the estimation of a joint model, the information shared by the two submodels can make the inferential process expensive and computationally unstable when the dimension of random effects increases. In order to make the inference process as efficient as possible, we propose a two-step approach. The first step consists of adjusting a longitudinal model without taking into account the survival process. In the second step we incorporate, as covariates, the estimates of the random effects from the longitudinal model into the survival process. It is well known that this approach produces bias estimation of the parameters for the survival model. Therefore, we assume that these covariates are measured with error to correct for this bias. Hence, we implement the simulation-extrapolation (SIMEX) algorithm to estimate the parameters of the survival model. We present results from simulations that compare the computational time and the robustness of our proposal with traditional approaches for joint models.

Keywords: Bias, joint models, SIMEX.

Multiple comparisons in poultry production trials

Palacios Luciano[1], *Hugo Delfino*[2], *O. Susana Filippini*[3], *Carla Martínez*[4]

[1]lucianofepalacios@hotmail.com, Department of Statistics, University of Luján.
[2]h_delfino@yahoo.com.ar.
[3]filippin@agro.uba.ar, Faculty of Agronomy, University of Buenos Aires.
[4] carla.r.martinez@gmail.com, Department of Statistics, University of Luján

Procedures of multiple comparisons allow us to detect differences among treatment means, which is the aim of many experiments on animals: the detection of significant relations. The use of significant tests is widely used. However, the benefits of these tests have been overvalued whenever they were considered an only indicator of the truthfulness of a hypothesis of "quantification" of the differences, the consideration of the error rates and the power of the test were omitted. Thus, conclusions tend to be wrong with consequence on animal performance, their health and welfare. The aim of this work is to develop concepts about the power of the tests: the error rates and quantification of testing, getting, in this way, of different tests ranking. As an example, broilers were used to estimate the effect of the addition of enzymes on diets of corn or triticale. Eight different treatments were used. All of them with three replicates: without cellulase enzymes, with addition of cellulase, protease and phytase, β-glucanase and with the only addition of phytase. The productive response measured were the live weight, accumulated consumption and food conversion of feeding. Several multiple comparison techniques are presented: Less Significant Differences test, Tukey test, Duncan test, Bonferroni test, Waller Duncan test, Student -Newman-Keuls test, etc. The procedures are grouped according to error rate, power and distributions used, to evaluate the most appropriate for this example.

Keywords: Multiple comparison, statistical power analysis, error rate.

Upscaling plant traits with Bayesian inference approaches

J. Martínez-Minaya[1], *A. Moreno-Martínez*[2], *D. Conesa*[1], *G. Camps-Valls*[2]

[1]joaquin.martinez-minaya@uv.es, david.v.conesa@uv.es, Department of Statistics and O.R., Universitat de València

[2]alvaro.moreno@uv.es, gustau.camps@uv.es, Image Processing Laboratory, Universitat de València

Plant traits are measurable features at the organismal level such as morphological, anatomical, physiological or phenological characteristics. They are essentially any attribute that influences the establishment, survival, or fitness of the considered individual. A wide diversity of plant traits are routinely measured in many sparse locations across the globe, and spread during many years. Despite the availability of these databases, little efforts have been done towards obtaining spatially explicit maps of selected plant traits. Previous work introduced random forests for spatialization using optical remote sensing data from operational satellites and climate data for the up-scaling of in-situ measured plant traits.

Nevertheless, the approach did not include spatial information or a solid Bayesian treatment of uncertainty. This work explores two alternative Bayesian approaches, the first one being a nonparametric regression model that spatializes the key plant traits (specific leaf area, leaf dry matter content, and leaf nitrogen concentration). We focus on Gaussian processes that can provide sensible confidence intervals for the predictions. The second approach is the usual in Species distribution models that incorporates the spatial structure of the data. In this case, posterior inference and prediction is done via the integrated nested Laplace approximation (INLA).

The methods are thoroughly compared in terms of accuracy, robustness and efficiency. Along with the estimated global maps of plant traits, we provide associated uncertainty estimates derived from the regression models, and a selection of the best model with the most important covariates. The proposed techniques applied allow attribution of information gain to data input and thus provide the opportunity to understand trait-environment relationships at the plant and ecosystem scale. Moreover, this comparison provides also a bridge between two disciplines: machine learning and statistics. Finally, the new data products can complement existing large-scale observations of the land surface and could anticipate substantial contributions to advances in quantifying, understanding and prediction of the Earth system.

Keywords: Plant traits, remote sensing, Bayesian statistics.

AMS: 62P12, 62G08, 62J12

ORdensity: an R package to identify differentially expressed genes

J.M. Martínez-Otzeta[1], *I. Irigoien*[2], *C. Arenas*[3], *B. Sierra*[4]

[1]josemaria.martinezo@ehu.eus, Department of Computer Sciences, University of Basque Country (UPV/EHU)

[2]itziar.irigoien@ehu.eus, Department of Computer Sciences, University of Basque Country (UPV/EHU)

[3]carenas@ub.edu, Department of Genetics, Microbiology and Statistics: Section of Statistics, University of Barcelona (UB)

[4]b.sierra@ehu.eus, Department of Computer Sciences, University of Basque Country (UPV/EHU)

An important issue in microarray data is to select, from thousands of genes, a small number of informative differentially expressed (DE) genes which may be key elements for a disease. We present an R package, called `ORdensity`, that implements a recently methodology developed in order to identify such genes. As output, it provides three measures related to the concepts of outlier and density of false positives in a neighbourhood, which allow to identify the DE genes with high classification accuracy.

An intrinsic computational issue in this context is the enormous amount of data that is involved. We alleviated this problem by means of a parallel implementation of the bootstrap procedure. The experimentation we carried out shows, on the one hand that for studies involving more than 5000 genes the paralellized implementation offers a clear improvement in runtime. On the other one, the experimentation also shows that for very large data sets (more than 15000 genes) the communication burden between the processes in the parallel implementation gets too big and there is no benefit in the runtime. Moreover memory allocation is still an issue we are working on.

As a summary, the new package offers an efficient implementation of a recent method to identify DE genes as well as a friendly and easy-to-use way for users not familiar with programming.

Keywords: Differentially expressed gene, outlier, R language

Fusing optical and microwave data using distribution regression for crop yield estimation

Anna Mateo-Sanchis[1], *José Adsuara*[1], *Maria Piles* [1], *Adrián Pérez-Suay*[1], *Gustau Camps-Valls*[1], *Jordi Muñoz-Marí*[1]

[1]Anna.mateo@uv.es, Image Processing Laboratory, University of Valencia

Remote sensing data provide a unique source of information to monitor crops in a spatially explicit way and temporally resolved. This is of outstanding relevance given the increase of world population and the ever-growing demand of food. Traditional remote sensing applications have exploited vegetation indices (VIs) to monitor the phenology of crops, and have vastly relied on summarizing the time series in a set of spatial and/or temporal descriptors. It is customary to summarize Earth observation (EO) time series with temporal metrics like the maximum peak or growing season, as well as to summarize all pixel-based observations within a area with the spatial average. Two temporal series can also be summarized using a sunergistic metric (e.g. the lag) or a Principal Component analysis.

We postulate here that summarizing is not always the best choice, and propose two nonlinear regression methods to account for all time and space observations that allows combining optical and microwave sensor observations. We illustrate the performance of the methods in two scenarios. First, we blend synergistically optical (MODIS EVI) and microwave (SMAP-VOD) data using full time series stacked at county level over the U.S. corn belt. Such data are then fed into a linear and kernel ridge regression to obtain county-based crop yield estimates. It is shown that the kernel regression outperforms the linear counterpart, and that the use of full time series from multisensor data improves the results obtained with common metrics and single sensors. The second experiment takes into account all goals simultaneously. In this case, we follow a distribution regression strategy that does not need to summarize the conduct of each county in an averaged time series. This machine learning method exploits higher-order relationship between all time series in a county, allows working with the native spatial resolution of each sensor. Results confirm the validity of the multisensor fusion and also the advantage of using distribution regression models with full time series for crop yield estimation.

Keywords: Remote sensing, machine learning, vegetation indices.

Understanding sediment provenance of soils at the archaeological site of Engaruka (Tanzania). A Bayesian approach.

Carlos J. Peña[1], *Carmen Armero*[2], *Agustín Pastor*[3], *Marco Lezzerini*[4], *Simon Chenery*[5], *Daryl Stump*[6], *Gianni Gallello*[7]

[1]carjape@alumni.uv.es, Department of Statistics and Operations research, University of València.
[2]carmen.armero@uv.es, Department of Statistics and Operations research, University of València.
[3]agustin.pastor@uv.es, Department of Analytical Chemistry, University of València
[4]marco.lezzerini@unipi.it, Department of Earth Sciences, University of Pisa
[5]srch@bgs.ac.uk, British Geological Survey, Environmental Science Centre
[6] daryl.stump@york.ac.uk, Department of Environment and Geography, University of York
[7] gianni.gallello@york.ac.uk, Department of Archaeology, University of York

The main aim of this work is to comprehend the source of alluvial deposits in the region of Engaruka - located on the Great Rift Valley of northern Tanzania - which are believed to have been transported by water through an ancient irrigation system for agricultural purposes. Data for the study consist of 162 observations from soil samples originated from three groups: *north field, south field* and *sources*. The latter group refers to the provenance of those soil samples from both fields, whose identification is our main interest in the study. Each observation contains information on the concentration (in *parts per million*) of 55 different chemical elements and compositions (in *percentages*) for 12 different compounds. The statistical analysis starts by performing a principal component analysis (PCA) for reducing the dimensionality of the set of chemical variables. Additionally, the behavior of the main PCA components is analysed by means of a Bayesian hierarchical model which accounts for the general properties of soil, sediment sources and fields as well as individual characteristics of the groups modeled via random effects.

Identification of factors associated with missing data in real world health care databases

Juan José Piñero de Armas[1]

[1]jjpinero@ucam.edu, Cátedra de Bioestadística y Big Data, Universidad Católica de Murcia

Databases that automatically records the medical activity in the health systems are growing relentlessly. This data has an enormous potential to advance biomedical and epidemiological research, but it comes with a set of challenging problems. One of them is the presence of large amounts of missing data in some variables. This is a problem that can bias the estimates and diminishing their precision. There are different methods to impute missing data that partly palliate the problem. But these methods rely on using good imputation models. It is therefore crucial to find out what other variables are key factors to explain the missingness. These factors could occur at different operational levels: related to the patient, the health staff, the health centers or to the system. Given the complexity of the data recoded at different levels the statistical models to explain missingness can become complex mixed multilevel regressions. We built models to analyse the factors determinant of missingness in other variables of interest (such as weight or number of cigarettes smoked) in a subset of primary care database extracted from of the public health service of Madrid. This data contains information collected over five years from almost a quarter of a million people. We developed three different approaches to deal with such a large amount of data: 1) Using the complete database, 2) Doing separate analyses in each health centre and pulling the results with a meta-analysis, 3) Doing separate analyses in random partitions of the database and then pulling the results with a meta-analysis. The results show that, for example, the OR of the variable weight being missing was 0.81 for women compared to men ($p < 10^{-16}$), 2.18 for foreigners compared to nationals ($p < 10^{-16}$). 0.94 for every 1 year increase in the Age ($p < 10^{-16}$). At the centre level, 1 unit increase in the doctor workload pressure index decreases has an OR of 0.85 ($p < 10^{-16}$) and 1 unit increase in nursing workload pressure has an OR of 1.18 ($p < 10^{-16}$). We also compared the results obtained with the three approaches. The most accurate results were obtained from the analysis of the whole database but it is extremely slow and often impossible to run in an average desktop computer. Fragmenting the analysis in randomized blocks provides good results and is much faster. By-center analysis is convenient as it does not require centers to share data but it might require a complex meta-egression if there are heterogeneity between centers.

Trilinear mixed model visualization for the analysis of the genotype, environment and management interaction in wheat

Victor Prieto[1], *Juan Burgueño*[2]

[1]vprieto@fagro.edu.uy, Departamento de Biometría, Estadística y Computación, Universidad de la República, Uruguay.

[2]j.burgueno@cgiar.org, Centro Internacional de Mejoramiento de Maíz y Trigo, CIMMYT, México.

In plant breeding programs, an extensive number of genotypes are evaluated in a wide range of environments including sites, years and crop management practices. This is due to the different genetic expression dependent on the environment, known commonly as genotype by environment interaction. To select superior genotypes (high yields, stability over the years, etc.) it is necessary to conduct series of multi-environments trials (MET) or combination of sites and years to evaluate the response to a particular environment of interest or target environment. The analysis of the data presents several complexities (multiplicity of factors, missing data, etc.) that the present work tried to address. From the extensive bibliography on the topic, emerge the linear-bilinear models as analysis strategies, being the model AMMI and GGE the most widespread. Its main use is regions and subregions delimitation based on biplot type graphics, for balanced data and two-way settings. Recently, the application of these models for higher order factorial designs has been discussed, considering the genotype and location, year and management factor, etc. Although these models have beneficial descriptive value, they are fixed-effect models, so the treatment of random factors is not possible. Mixed effects model approach does allow it, with the advantage of being able to use particular covariance structures to model relations between sites, years, and related genotypes, etc., as well as the possibility of handling unbalanced data within certain limits. It is of interest to evaluate the factor analytic (FA) mixed model and their graphic possibilities. The main purpose of this work was to evaluate the advantages of the use of mixed linear-bilinear and trilinear models in the analysis of the triple interaction genotype by environment by management.

MET experimental data consisted of a complete dataset of 35 genotypes of bread and durum wheat evaluated in eight years at a unique location (Obregon, México). The trait analyzed in this dataset was grain yield and the trials were in randomized complete block design with 3 replicates. Four management systems were evaluated in each year: a combination of two levels of Tillage management and two Irrigation management. Different linear mixed models were fitted comprising effects of genotype, management, year and their interactions.

From the results obtained, it was possible to obtain stability indicators for the genotype-year-management interactions without having to collapse or average factors, allowing the evaluation of the genetic materials from the perspective of management and year. Mixed bi-triplot were developed visualizing the three main factors under study. This methodology aims to be of valuable use in genetic improvement programs where data complexity is expected.

Keywords: genotype by environment interaction, mixed model, factor analytical model.

Predictive modeling of digital pathology images with integration of computer extracted tissue biomarkers and genomics data

Ferran Reverter[1], *Esteban Vegas*[2]

[1]freverter@ub.edu, Department of Genetics, Microbiology and Statistics, University of Barcelona

[2]evegas@ub.edu, Department of Genetics, Microbiology and Statistics, University of Barcelona

Histopathological high-resolution microscopic images (whole slide image, WSI) of healthy or diseased tissue samples that have been sectioned and stained are essential for identifying and characterizing complex phenotypes. Pathologists study tissues using histological imaging techniques for scientific research on cellular morphology and tissue structure and for medical practice. The Cancer Genome Atlas (TCGA) contains over 10,000 WSIs from various cancer types, this database may serve as potential training data for various tasks. TCGA also provide genomic profiles, which could be used to investigate relationships between genomics and morphology.

Considerable research has been done in the computational analysis of pathological image data to develop automatic cancer diagnoses. Earlier techniques generally involved the extraction of predefined morphological, color and textural image features from the histological images. Many approaches based on hand-made features are not implicitly prepared to manipulate and distill large data sets into classifiers in an efficient manner. On the other hand, Deep Learning (DL) approaches work well in these circumstances. DL iteratively upgrades the representations learned from the underlying data in order to discriminate class patterns.

In this study we automatically extract image features using convolutional autoencoder (CAE). A CAE is an unsupervised DL method that produces a small set of numeric features characterizing each input image that allows the reconstruction of the input images with minimal loss. These image feature representations are intended to capture variance in the image as a whole, but we can also produce image features that are predictive of class labels, such as tumor versus healthy samples. Clustering these image signatures aggregates tumors into groups with cohesive morphologic characteristics. Then, clusters were inspected to find associations with transcriptional events. This methodology is demonstrated with an analysis of glioblastoma, using data from TCGA.

Keywords: Deep learning, convolutional autoencoder, whole slide image, TCGA.

AMS: 62H30, 62H35.

Comparing GWAS models for genetically correlated multi-environment data

Rueda Calderón A.[1], *Bruno C.*[2], *Balzarini, M.*

[1]angelicarueda@agro.unc.edu.ar, School of Agricultural Sciences, National University of Córdoba (UNC), National Scientific and Technical Research Council (CONICET)

[2]cebruno@agro.unc.edu.ar, School of Agricultural Sciences, National University of Córdoba (UNC), National Scientific and Technical Research Council (CONICET)

[3]mbalzarini@agro.unc.edu.ar, School of Agricultural Sciences, National University of Córdoba (UNC), National Scientific and Technical Research Council (CONICET).

Within a context of highly available molecular marker information, and with the omnipresence of multi-environment trials (MET) to assess genotype (G) performance in different environments (E), genome-wide association studies (GWAS) require models to handle multi-environment data. When the genotype (G) effects are assessed from MET, the effects of genotype by environment interaction (G×E) can impact results. Besides, prediction of the G and marker effects can be also affected by genetic correlations. The aim of this study was to compare the accuracy of different statistical approaches that use genome-wide genetic and pedigree information to account for genetic correlation in GWAS from MET. Data comprises 599 wheat lines genotyped with 1279 molecular markers assessed with 3 replicates in 4 environments. The compared models were: M1-Single-environment model fitted by adding pedigree information to account for correlations among lines; M2-Single-environment model with correlations among lines estimated from molecular similarity; M3-MET model involving pedigree information; and M4-MET model involving molecular similarity. For each model, variance components, marker scores and GBLUP were obtained. The magnitude of the genetic variability estimates depended on the information used to model genetic correlations. Selected genotypes were similar in single-and multi-environment models. However, MET models are preferred because is possible to quantify G×E and to obtain fitting criteria to select the best model for the underlying genetic correlations.

Keywords: GBLUP, marker scores, G×E interaction, pedigree, molecular similarity.

Metals and plasma metabolomics from a population-based study

Francisco Sánchez-Sáez[1],*Maria Grau-Pérez*[1,2], *Daniel Monleón Salvado*[1,3], *Maria Téllez-Plaza*[1,4,5]

[1]Institute for Biomedical Research INCLIVA, Valencia.
[2]Universidad Autónoma de Madrid, Madrid.
[3]University of Valencia, Valencia.
[4]National Center for Epidemiology, Carlos III Health Institutes, Madrid.
[5]Johns Hopkins University, Baltimore

Introduction: Metabolite profiling has successfully been applied to identify biomarkers of health issues. However, very few and small studies to date have evaluated the potential association of metals exposure with circulating metabolites levels. We analyzed the relationship of 13 concentration of metals and 49 measures of metabolites in 1166 participants from the Hortega Study, a population-based study from Spain.

Methods: Concentrations of 10 urine (u.) metals and 3 plasma (p.) metals were measured. Urine concentrations of cobalt, copper, molybdenum, zinc, antimony, arsenic, barium, cadmium, chromium and vanadium were measured using ICPMS at the University of Huelva, Spain. Urine dilution is accounted by dividing all metal concentrations by urine creatinine levels. Plasma concentrations of copper, selenium and zinc were measured by atomic absorption spectrometry with graphite furnace at Cerba Int. Lab. Ltd. The quantification of the 49 metabolites were performed as follows: 32 from serum metabolomic profile assessed by a proton NMR spectra and 17 lipoprotein subclasses obtained from a Liposcale test. The association between metals and metabolites was evaluated in three ways: (1) by single-metal lineal regression models; (2) by multi-metal lineal regression models; and (3) evaluating the association of metal mixtures using principal components (PC) analysis.

Results: 25 metabolites were significatively associated with several metals in the single-metal regression models. In particular, 5 metabolites were associated with p.copper, 7 with p.selenium, 3 with p.zinc, 3 with both p.copper and p.selenium and 7 with both p.selenium and p.zinc. Almost all of these associations remained significant in multi-metal models. In PC analysis, the 13 metals were grouped into 2 PC, reflecting the 10 urine metals together (PC1) and the 3 plasma metals together (PC2). Consistently, PC2 was significatively associated with the same metabolites as the 3 plasma individual metals.

Conclusion: Plasma levels of copper, selenium and zinc were especially relevant, as they showed a significant association with several metabolites. Regarding the direction of the associations, we observed that metabolites belonging to fatty acids, cholesterol and triglycerides classes were positively associated with plasma metals. These results are in line with previous research, as high concentrations of these metabolites are traditionally considered as risk factors for cardiovascular related-outcomes and diabetes, health issues for whose some of these metals have also been associated. On the other hand, metabolites belonging to the amino acids, which are traditionally considered as protector factor for the aforementioned diseases, were negatively associated with these metals. More studies are needed to confirm these results. Public health interventions that prevent metal exposure in the population may be needed.

Keywords: Metals, metabolomics, Principal Components.

AMS: 97K80

A comparative of designs for Michaelis-Menten models with high-substrate inhibition

Santos-Martín, M.T[1], *Mariñas-Collado, I*[2], *Rivas-López, M.J.*[3], *Rodríguez-Díaz, J.M*[4]

[1]maysam@usal.es, Department of Statistics, University of Salamanca.

[2]irenemc@usal.es, Department of Statistics, University of Salamanca.

[3]chusrl@usal.es, Department of Statistics, University of Salamanca.

[4]juanmrod@usal.es, Department of Statistics, University of Salamanca.

Kinetic reactions usually follow a Michaelis-Menten (MM) model, extensively used in biochemistry. The problem is that sometimes, while the MM equation is obeyed at lower substrate concentrations, above a critical amount of substrate, the data deviate significantly from the anticipated behaviour, showing a lower than expected value for the measured velocity. There are two variations of the model to account for these deviations from the MM kinetics. We present a comparison between the most commonly used designs (such as uniform, arithmetic and parabolic) and D-optimal designs for these two models in an application to the semicarbazide-sensitive amine oxidise (SSAO) activity towards benzylamine, based on the designs efficiency.

Keywords: D-optimality, Michaelis-Menten model, inhibition.

AMS: 97K80

Performance of sampling strategies for delimiting Xylella fastidiosa infection: in Alicante.

M. Sesé[1], *E. Lázaro*[2], *D. Conesa*[3], *A. López-Quilez*[2], *V. Dalmau*[3], *A. Ferrer*[3], *A. Vicent*[1]

[1]misefi@alumni.uv.es, Centre de Protecció Vegetal i Biotecnologia, Institut Valencià d'Investigacions Agràries.

[2]Departament d'Estadística e Investigació Operativa, Universitat de València.

[3]Servei de Sanitat Vegetal, Conselleria d'Agricultura, Medi Ambient, Canvi Climàtic i Desenvolupament Rural.

Xylella fastidiosa is a phytopathogenic bacterium whose presence has been confirmed in several countries in the European Union (EU). At that regard, EU has implemented multiple protective and emergency measures against its introduction and its further spread. These legal measures consider among other actions, the implementation of a delimiting survey whether the presence of the bacterium had been confirmed. The objective of a delimiting survey consists on demarcating the geographic extent of the disease as well as the application of the eradication or containment measures, depending on the case. One of the affected regions in the UE is Alicante, for this reason it is currently subjected to an intensive surveillance program. Since the first detection was confirmed, more than 100.000 has. have been demarcated and surveyed with around 20000 samples have been analyzed. Under this framework, the aim of this work is to propose an alternative delimiting survey strategy to improve the effectiveness of the current one. Based on surveillance data collected during the 2018 several sampling strategies have been simulated to find an optimum sampling intensity. The effectiveness of the sampling strategies has been assessed by means of a comparison of the reference data in terms of delimitation efficacy and disease prevalence estimates.

Keywords: Delimiting surveys, spatial sampling, simulation-optimization.

Anylisis of soil moisture from descriptive and explanatory methods multivariate

Zabala, Stella Maris[1], *Salgado Héctor* [2], *Filippini O. Susana* [3]

[1]stezab@hotmail.com, Department of Statistics, University of Luján.

[2]hsalgado@agro.uba.ar, Faculty of Agronomy, University of Buenos Aires.

[3]sfilippini@unlu.edu.ar, Department of Statistics, University of Luján. Faculty of Agronomy, University of Buenos Aires

The temporary monitoring of the water contained in the soil is of great interest both for agricultural producers, as well as for professionals and researchers from multiple disciplines. The measurement of soil moisture (HS) is difficult, given its high spatial and temporal variability. The incidence of different climatic and soil variables on the volumetric water content (CVA) available for the vegetation for its prediction is analyzed. The experience is carried out in an area of intensive agriculture, in the center of the Province of Buenos Aires, Argentina, based on the complementation of hydro-meteorological data and land campaigns. Field measurements were carried out using dielectric probes and by the gravimetric standard method. The analysis of the data was carried out in two stages. In the first, the dimensionality of the database was reduced. Subsequently, a multiple linear regression model was applied, considering the previously obtained factors as predictor variables to analyze the relationships and variability of the studied system. The obtained results show that, the reduction of the dimensionality of the information matrix obtained helped to find a general linear function that will allow to predict the volumetric content of water in soils in the humid Pampas region in a simple way. The experience is encouraging to continue with studies and adjustments, tending to achieve a daily monitoring of the HS in the zonal agricultural plot.

Keywords: Soil moisture monitoring, climatic variables, multivariate methods.

AMS: Statistical methods in Agriculture.

Index