CEBMADRID 2022

Book of Abstracts

XVIII Congreso de Biometría CEBMADRID

Madrid, May 25-27, 2022

XVIII Congreso de Biometría CEBMADRID
25, 26 y 27 mayo 2022, Madrid

Editors:

Stefano Cabras	(ORCID: 0000-0001-6690-8378)
Ignacio Cascos	(ORCID: 0000-0001-8461-746X)
María Eugenia Castellanos	(ORCID: 0000-0001-7920-2307)
María Durbán	(ORCID: 0000-0002-4272-7895)

Book of Abstracts
XVIII Congreso de Biometría
CEBMADRID

SEB
SOCIEDAD ESPAÑOLA
DE BIOESTADÍSTICA

IBS

uc3m | Universidad
Carlos III
de Madrid

Universidad
Rey Juan Carlos

fundación
MUTUALIDAD
ABOGACÍA

ISBN: 978-84-16829-70-5
XVIII Congreso de Biometría CEBMADRID
Madrid, May 25-27, 2022.

Editors: Stefano Cabras, Ignacio Cascos, María Eugenia Castellanos, María Durbán

Electronic version available at e-Archivo: `http://hdl.handle.net/10016/34695`

Scientific Programm Committee

- María Eugenia Castellanos (Chair)
  Universidad Rey Juan Carlos
- Carmen Armero
  Universitat de València
- Inmaculada Arostegui
  Universidad del País Vasco
- María Durbán
  Universidad Carlos III
- Virgilio Gómez Rubio
  Universidad de Castilla-La Mancha
- Ana Justel
  Universidad Autónoma de Madrid
- Dae-Jin Lee
  Basque Center for Applied Mathematics
- Vicente Núñez-Antón
  Universidad del País Vasco
- Nuria Porta
  Institute of Cancer Research, London
- Pere Puig
  Universidad Autónoma de Barcelona
- Rebeca Ramis
  Instituto de Salud Carlos III
- Mar Rodríguez Girondo
  Leiden University Medical Center
- Bruno de Sousa
  University of Coimbra. CINEICC
- Aurelio Tobias
  Consejo Superior de Investigaciones Científicas (CSIC)

Preface

The CEB conference is held in Madrid, May 25-27, 2022. This is the 18th meeting of the Spanish Biostatistics Society (`http://www.biometricsociety.net/`) last earlier meetings held in Valencia (2019) and Sevilla (2017).

The principal objectives of CEB is to facilitate the exchange of recent research developments in Biostatistics, to provide opportunities for new researchers, and to establish new collaborations and partnerships.

Welcome to Madrid. Enjoy the city and surroundings and have a great conference.

Stefano Cabras, Ignacio Cascos, María Eugenia Castellanos, María Durbán

Madrid, May 2022

Contents

Plenary sessions

First session young researchers

Chairperson: Inmaculada Arostegui, Universidad del Pais Vasco (Spain)

Parallel session: Epidemiological studies

Chairperson: Miguel Angel Martínez, Universitat de Valencia (Spain)

Parallel session: Medical studies 1

Chairperson: Xabier Barber, Universidad Miguel Hernández (Spain)

Parallel session: Omic data analyses

Chairperson: Stefano Cabras, Universidad Carlos III de Madrid (Spain)

Parallel session: Covid-19

Chairperson: Paloma Botella, Generalitat Valenciana (Spain)

Parallel session: Biostatistical methods

Chairperson: Dae-Jin Lee, Basque Center for Applied Mathematics (Spain)

Parallel session: Software

Chairperson: Natalia Vilor-Tejedor Beta Brain Research Center (Spain)

Poster session

Chairperson: M. E. Castellanos, URJC (Spain) and V. Gomez Rubio, UCLM (Spain)

Parallel session: Sport

Chairperson: Irantzu Barrio, Universidad del País Vasco (Spain)

Parallel session: Ageing

Chairperson: Maria Durban, Universidad Carlos III de Madrid (Spain)

Parallel session: Regional IBS

Chairperson: Malu Calle, Universitat de Vic (Spain)

Parallel session: Bayesian analysis in medical studies

Chairperson: Antonio López-Quilez, Universitat de Valencia (Spain)

Parallel session: Medical studies 2

Chairperson: Christian Tebé, IDIBELL, Unitat de Bioestadística (Spain)

Second session young researchers

Chairperson: Ana Justel, Universidad Autonoma de Madrid (Spain)

Parallel session: Survival analysis

Chairperson: Klaus Langohr, Universitat Politecnica de Catalunya (Spain)

Parallel session: Bio-Bayes modelling

Chairperson: M. Eugenia Castellanos, Universidad Rey Juan Carlos (Spain)

Plenary sessions

Plenary session 1

Chairperson: Anabel Forte, Universitat de Valencia (Spain)

How to correct for baseline covariates in longitudinal clinical trials?

Geert Verbeke, Ku Leuven University (Belgium)

Abstract In clinical trials, mixed models are becoming more popular for the analysis of longitudinal data. The main motivation is often expected dropout which can easily be handled through the analysis of the longitudinal trajectories. In many situations, analyses are corrected for baseline covariates such as study site or stratification variables. Key questions are then how to perform a longitudinal analysis correcting for baseline covariates, and how sensitive are the results with respect to choices made and models used ? In this presentation, we will first present and compare a number of techniques available to correct for baseline covariates within the context of the linear mixed model for continuous outcomes. Second, we will study the sensitivity of the various techniques in case the baseline correction is based on a wrong model or does not include important covariates. Finally, our findings will be used to formulate some general guidelines relevant in a clinical trial context. All findings and results will be illustrated extensively using data from a real clinical trial.

Plenary session 2

Chairperson: Guadalupe Gomez, Universitat Politecnica de Catalunya (Spain)

Modeling the COVID-19 epidemic in Belgium to inform policy makers

Christel Faes, Hasselt University (Belgium)

Abstract Belgium has been hit particularly hard by the coronavirus placing the country near the top in international rankings when looking at the number of confirmed cases per 100,000 and the number of deaths per million. Belgium accounted for more than half a million confirmed cases and over 17,000 SARS-CoV-2 confirmed and suspected deaths in 2020. Belgium's location at the centre of Europe, high international mobility, high population density, high average household size and an older population structure combined with a relatively high mixing behaviour increases transmission potential. Short-term predictions were used to help local and national governments in decision-making on interventions during the outbreak and preserving the hospital capacity. Information on local mobility, absenteeism, testing strategy and GP consultations are used in the prediction model, using distributed lag non-linear models. Spatio-temporal trends are tracked to raise alarms when growth rate in hospitalizations and cases change. Mathematical modelling was used to inform policy makers on the possible impact of restriction measures. Some highlights of the modelling exercises will be presented.

Plenary session 3

Chairperson: David Conesa, Universitat de Valencia (Spain)

Sloppy models: unveiling parameter uncertainty in mathematical models

Kerrie Mengersen, Queensland University of Technology (Australia)

Abstract In this presentation, I will discuss a Bayesian approach to assessing the sensitivity of model outputs to changes in parameter values in mathematical models, constrained by the combination of prior beliefs and data. The approach identifies stiff parameter combinations that strongly affect the quality of the model-data fit while simultaneously revealing which of these key parameter combinations are informed primarily from the data or are also substantively influenced by the priors. These stiff parameter combinations can uncover controlling mechanisms underlying the system being modeled and guide future experiments for improved parameter inference.

The focus of the discussion will be on the very common context in complex systems where the amount and quality of data are low compared to the number of model parameters to be collectively estimated. The approach will be illustrated with applications in biochemistry, ecology, and cardiac electrophysiology.

This work is joint with Gloria Monsalve-Bravo (lead author), Brodie Lawson, Christopher Drovandi, Kevin Burrage, Kevin Brown, Christopher Baker, Sarah Vollert, Eve McDonald-Madden and Matthew Adams.

The full paper is available as an arXiv preprint arXiv:2203.15184

Plenary session 4

Chairperson: Carmen Armero, Universitat de Valencia (Spain)

I've been irradiated!! What is the total amount of radiation I've received?

Pere Puig, Universidad Autónoma de Barcelona (Spain)

Abstract In the event of a radiation accident, biological dosimetry is critical for determining the radiation dose received by an exposed individual in a timely way. The dose is estimated by calculating the amount of damage caused by radiation at the cellular level, such as by counting the number of chromosome aberrations like dicentrics micronuclei, or translocations. The theory of count data distributions is critical for achieving this goal. In this talk, we will introduce the standard statistical methodology for dose estimation described in the International Atomic Energy Agency's manual (IAEA, 2011), as well as summarise recent research led by our team. We will present models based on compound Poisson processes that are suitable for describing high-LET radiation exposures such as those seen in the Fukushima accident, zero-inflated and mixed Poisson models for partial and heterogeneous exposures, and weighted Poisson models for integrating low and high doses.

First session young researchers

Chairperson:

Inmaculada Arostegui, Universidad del Pais Vasco (Spain)

INLAMSM: Adjusting multivariate lattice models with R and INLA

Francisco Palmí Perales[1*], Virgilio Gómez Rubio[2] and Miguel Ángel Martínez Beneito[3]

[1] Departamento de Economía Aplicada, Universitat de València
[2] Departamento de Matemáticas, Universidad de Castilla-La Mancha
[3] Departamento de Estadística e Investigación Operativa, Universitat de València
[*] Corresponding author

Abstract

Fitting multivariate spatial models becomes faster and easier with INLAMSM. This R package provides a collection of multivariate spatial models for analysing lattice data. The implemented models, which include different structures to model the variables' spatial variantion and the between-variables variability, can be used (with R-INLA) for performing Bayesian inference. Two different datasets have been used to exemplify the use of the package.

Keywords: INLAMSM, lattice data, R package

1. Introduction

The Integreted Nested Laplace Approximation [2] methodology focuses on estimating the posterior marginals of the model parameters instead of their joint posterior distribution. INLA is implemented in the R-INLA package for the R programming language. These package implements several likelihoods, priors and latent effects that can be used to build models. However, multivariate spatial models can not be fitted (only) with R-INLA. The INLAMSM package adds a number of multivariate latent effects that implement classic multivariate spatial models for areal data. By fitting these models with INLA, instead of Markov chain Monte Carlo algorithms (MCMC), computing times should be reduced.

2. Multivariate spatial context

The different functions implemented in INLAMSM correspond to different multivariate spatial latent effects. Although, these functions differ in structure, complexity and number of hyperparameters, all of them are defined in a multivariate spatial context in which $i = 1, \ldots, I$ represents the spatial areas and $k = 1, \ldots, K$ is used to index the variables measured in region i.

Random effects can be represented using a matrix Θ with entries $\theta_{ik},\ i = 1, \ldots, I,\ k = 1, \ldots, K$. Hence, the k-th column of Θ represent the spatial random effects associated to variable k.

There are 5 different structures implemented in INLAMSM: independent intrinsic MCAR, independent proper MCAR, improper MCAR, proper MCAR i M-model. The last one is the one defined in [1]. This article details an unifying modelling framework for multivariate disease mapping when the number of diseases is potentially large. In this context, the number of the observed cases of k-th disease in the i-th spatial area, Y_{ik}, is modeled as a Poisson random variable:

$$Y_{ik} \sim Po(E_{ik} \cdot R_{ik})$$

where E_{ik} and R_{ik} represent the expected cases and the relative risk for the i-th spatial area and the k-th disease, respectively. Then, the logarithm of the relative risk is the sum of two terms:

$$\log(R_{ik}) = a_k + \theta_{ik}$$

where a_k is the intercept of the k-th disease and θ_{ik} is the term that models the spatial variability. Note that other covariates can be included in the linear predictor on the right hand side of the previous equation.

3. M-model

In the case of the M-model, the spatial effects are linear combinations of proper CAR spatial effects. In particular, we will consider K underlying proper CAR spatial effects defined by

$$\phi_k \sim N\left(0, (D - \alpha_k W)^{-1}\right), \ k = 1, \ldots, K.$$

Here, ϕ_k is a vector of length I.

The value of the spatial random effect $\Theta_{\cdot j}$ for variable j is defined as

$$\Theta_{\cdot j} = \phi_1 m_{1j} + \ldots \phi_J m_{Jj}.$$

Hence, matrix M with entries m_{ij} defines the loadings of the different underlying CAR spatial effects for each disease or variable.

The distribution of these random effects is given by

$$vec(\Theta) \sim N\left(0, (M^\top \otimes I) diag((\Sigma_w)_1, \ldots, (\Sigma_w)_K)(M \otimes I)\right).$$

Here, matrices $(\Sigma_w)_k$ are the variance matrices of the K underlying proper CAR spatial effects, i.e.,

$$(\Sigma_w)_k = (D - \alpha_k W)^{-1}, \ k = 1, \ldots, K.$$

Additionally, [1] show that for the separable case, with $\alpha_1 = \ldots = \alpha_K$, the between-variables variance matrix is $M^\top M$. The chosen prior of this model is on $M^\top M$, and it follows a Wishart with parameters K and τI. In this case τ is a fixed precision which is set to 0.001. Finally, the hyperparameters of this model are the K autocorrelation parameters (conveniently transformed) followed by the columns of matrix M, for which no transformation is required.

4. Mortality in Comunidad Valenciana

One of the two examples analysed in this work is based on simulated data of the mortality by cirrhosis, lung and oral cancer in Comunidad Valenciana (Spain). This dataset is available at `http://github.com/MigueBeneito/DisMapBook` and it has been generated to mimic the spatial pattern of the real data, that cannot be provided due to confidentiality constraints.

Specifically, the number of deaths by these three causes are available at the municipality level in Comunidad Valencia (Spain), as well as the expected number of cases that have been computed using internal standardization are available. Hence, the aim now is to estimate the spatial pattern of the different diseases as well as their possible correlations using (in this example) three different functions of `INLAMSM`: intrinsic MCAR, roper MCAR and the M-Model.

The maps in Figure 1 show the different posterior means of the relative risks from the models fitted for the different causes of death. Similar point estimates of the relative risks are produced by the three different models. The INLAMSM package also includes a few functions to transform the marginals and summary statistics of the model hyperparameters in the internal scale into the original scale in the model.

5. Discussion

The INLAMSM package builds on top of the INLA package and implements a number of multivariate spatial latent effects. Hence, this package allows an easy and simple definition of these multivariate effects to be used within a formula term to fit multivariate spatial models to lattice data.

This work has been published in an open acces journal [**?**], therefore, more details can be checked in `https://www.jstatsoft.org/article/view/v098i02`

6. Acknowledgments

This work has been supported by grants PPIC-2014-001-P and SBPLY/17/180501/000491, funded by Consejería de Educación, Cultura y Deportes (Junta de Comunidades de Castilla-La Mancha, Spain) and FEDER, grant MTM2016-77501-P, funded by Ministerio de Economía y Competitividad (Spain), grant PID2019-106341GB-I00 from Ministerio de Ciencia e Innovación (Spain) and a grant to support research groups by the University of Castilla-La Mancha (Spain). F. Palmí-Perales has been supported by a Ph.D. scholarship awarded by the University of Castilla-La Mancha (Spain).

7. Bibliography

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

[2] Palmí-Perales, F., Gómez-Rubio, V., and Martinez-Beneito, M. A. (2021). Bayesian Multivariate Spatial Models for Lattice Data with INLA. *Journal of Statistical Software*, 98(2), 1â29.

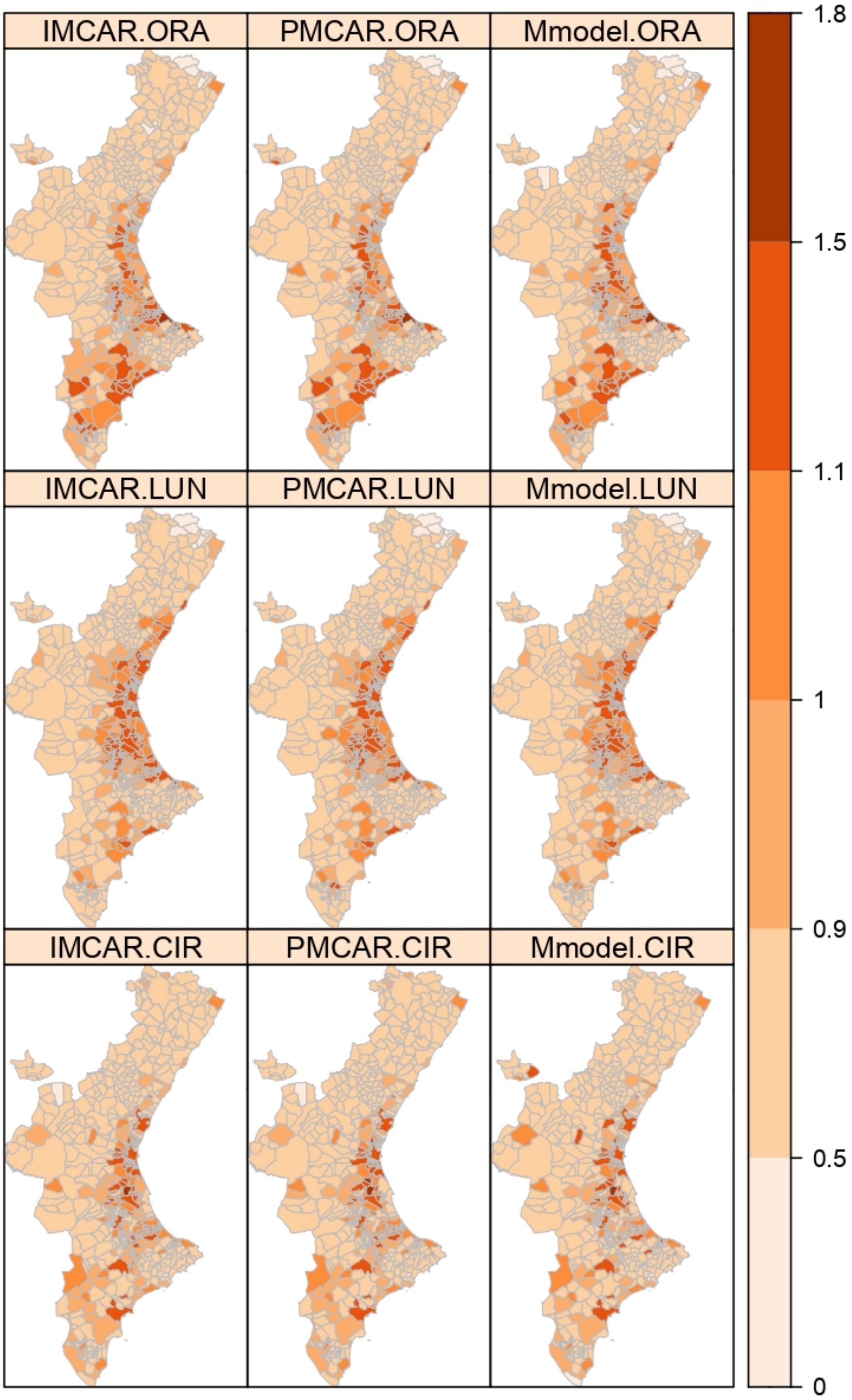


Figure 1: Posterior means of the relative risks of cirrhosis, lung cancer and oral cancer in Comunidad Valenciana (Spain).

[3] Rue H., Martino S., and 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.

Constrained smoothing and out-of-range prediction using $P-$splines: an application to COVID-19 evolution

Manuel Navarro García[1], Vanesa Guerrero Lozano[2], María Luz Durbán Reguera[3]

[1]mannavar@est-econ.uc3m.es, Department of Statistics, University Carlos III of Madrid, and Komorebi AI Technologies

[2]vanesa.guerrero@uc3m.es, Department of Statistics, University Carlos III of Madrid

[3]mdurban@est-econ.uc3m.es, Department of Statistics, University Carlos III of Madrid

Abstract

This work addresses the problem of constrained estimation of smooth curves and out-of-range prediction combining the statistical framework of $P-$splines [5] with conic optimization [2]. The methodologies proposed in this paper are illustrated using a real data set about of the evolution of the COVID-19 pandemic.

Keywords: P-splines, Conic optimization, Prediction

1. Introduction

Many evolution indicators of the COVID-19 pandemic, such as the number of daily positive cases or fatalities in a region, are non-negative responses. Therefore, any smoothing technique applied to this kind of data must be able to incorporate this requirement, or misleading estimations may be obtained as it is shown in Figure 2a. Furthermore, simulating different constrained short-term prediction scenarios, which include expert knowledge, is a challenge which has not been properly addressed so far in the $P-$splines framework.

In this work, we tackle the problem of estimating smooth curves that satisfy requirements about their sign and shape. To do so, statistical and mathematical optimization modeling paradigms merge to propose a $P-$splines constrained estimation problem by means of a conic optimization model. The proposed methodologies are applied to COVID-19 data, yielding coherent estimations of the daily number of infected people, contrary to its unconstrained version shown in Figure 2a. Moreover, expert knowledge can be incorporated into this framework to simulate different short-term prediction scenarios. We illustrate the flexibility of our approach by taking information about the growth rates in the first COVID-19 wave and incorporating them to carry out predictions in the second wave period. Furthermore, an open source Python library, `cpsplines`, is developed which contains the implementations of the proposed methodologies.

2. Constrained $P-$splines

Let us consider the univariate regression problem of estimating a continuous function $f : [x_1, x_n] \subset \mathbb{R} \longrightarrow \mathbb{R}$ such that

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where x_i and y_i refer to the $i-$th observation of the continuous covariate X and the response variable Y to be predicted, respectively. For this estimation, $P-$splines [5] were used, which consist of a basis function approach using splines for regression, together with a penalization term. Many different spline basis can be used, and we chose cubic $B-$splines [4] for this work.

Estimating the function f in model (1) using $P-$splines does not ensure that the estimated curve satisfies non-negativity requirements. Taking into account that the approximation of f in (1) using a $B-$spline basis is defined by cubic piecewise polynomials, we use the characterization of non-negative univariate polynomials in [1] to impose this condition. This result allows us to incorporate the non-negativity requirement into the $P-$splines framework yielding, for appropriate choices of the vector $\mathbf{c}$ and matrices $\mathbf{C}_q$, $\mathbf{H}_\ell$ (arisen from the result in [1]) and $\mathbf{P}$ (containing the basis and penalty matrices), the following conic optimization model:

$$\begin{array}{ll} \underset{u,\ \boldsymbol{\theta},\ \mathbf{Z}_q}{\text{minimize}} & u - \mathbf{c}^\top \boldsymbol{\theta} \\ \text{subject to} & \langle \mathbf{H}_\ell, \mathbf{Z}_q \rangle_F = (\mathbf{C}_q \boldsymbol{\theta})_\ell, \quad q = 4, 5, \dots, k+3, \quad \ell = 1, 2, 3, 4, \\ & \langle \mathbf{H}_\ell, \mathbf{Z}_q \rangle_F = 0, \quad q = 4, 5, \dots, k+3, \quad \ell = 5, 6, 7, \\ & \left(\mathbf{P}^{1/2}\boldsymbol{\theta}, u, \frac{1}{2}\right) \in \mathcal{Q}_r^{k+5}, \\ & \mathbf{Z}_q \in \mathbb{S}^4, \quad q = 4, 5, \dots, k+3, \\ & \boldsymbol{\theta} \in \mathbb{R}^{k+3}, \end{array} \tag{2}$$

where $\langle \cdot, \cdot \rangle_F$ denotes the trace matrix inner product, $(\cdot)_\ell$ refers to the $\ell-$th component of a vector, k denotes the number of subintervals in which the data domain $[x_1, x_n]$ is split, $\boldsymbol{\theta}$ contains the $B-$spline basis coefficients, $\mathbf{Z}_q$ are decision variables belonging to the cone of positive semidefinite matrices of order 4, $\mathbb{S}^4$, and $\mathcal{Q}_r^{k+5}$ is the $(k+5)-$dimensional rotated second-order cone.

The theoretical result used to state problem (2) and the work in [3] also allows us to develop some extensions. First, to extend the previous methodologies to out-of-range prediction (forward and backwards), we follow the approach in [3], in which the values to be predicted are treated as missing data in the fitting procedure. Second, thanks to the regularity properties of the $B-$splines basis, the non-negativity requirement may be applied to higher-order derivatives of the function approximating f in (1). Therefore, these methodologies can be generalized to enforce shape constraints in the estimated curve such as monotonocity and curvature. And third, our approach can also be extended to the case in which multiple curves have to be simultaneously estimated for groups of observed data and we wish to enforce a relative position in the curves.

When multiple regression is considered instead, i.e.,

$$y_i = f(x_{i,1}, \dots, x_{i,m}) + \varepsilon_i, \quad i = 1, \dots, n, \tag{3}$$

where $x_{i,j}$ refers to the $i-$th observations of a continuous predictor X_j, $j = 1, \dots, m$, $P-$splines are generalized by taking a multivariate $B-$spline basis, which is defined as the tensor product of the elements of univariate $B-$splines bases. However, the previous non-negativity characterization does not hold for multivariate polynomials, and hence other approaches have to be considered. We opted by weakening the requirements by just imposing them over a finite set of curves that belong to the hypersurface. Our approach in this multidimensional setting does not guarantee that the shape conditions are satisfied in the whole domain in which the hypersurface is being estimated. Nevertheless, if these conditions are imposed at a large enough number of curves, the violation of such conditions becomes unlikely in part due to the demanded smoothness.

3. Case study: Coherent estimations and predictions in COVID-19 data

We illustrate our new methodological proposal with data of the number of daily infected people by COVID-19 in Aragón from February 6th 2020 to February 15th 2021 and labelled depending on the age group of the patient ($0-9, 10-19, \dots, 70-79, 80+$). With this data set, our aim is to present different short-term prediction scenarios (setting the forecasting horizon in 14 days ahead), ensuring that the estimations are coherent with the nature of the data.

Figure 1 shows the fit and forward prediction of the number of daily positive cases in Aragón (aggregated by ages) from February 6th to July 15th 2020 under different scenarios. Scenarios 2, 3 and 4 are obtained by incorporating the growth rate of the first wave (varying its scale and/or adding a constant lag to the points where this value is enforced) at the points to be predicted, while in Scenario 1 only non-negativity is imposed, both in the observed domain and in the forecasting region. The results shown in Figure 1 highlight that the methodologies we propose and the incorporation of these two parameters (the one that scaling the derivative and the one lagging the curve) are a simple but very flexible tool to simulate different plausible scenarios in an epidemiological context. Moreover, we observe in Figure 2 that the enforcement of the non-negative requirement is crucial for obtaining reasonable smooth curves, even when all the observed points are non-negative as in this case.

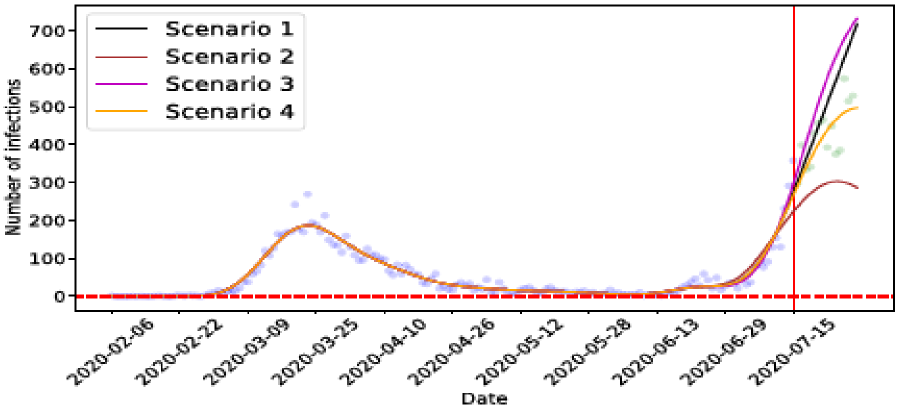
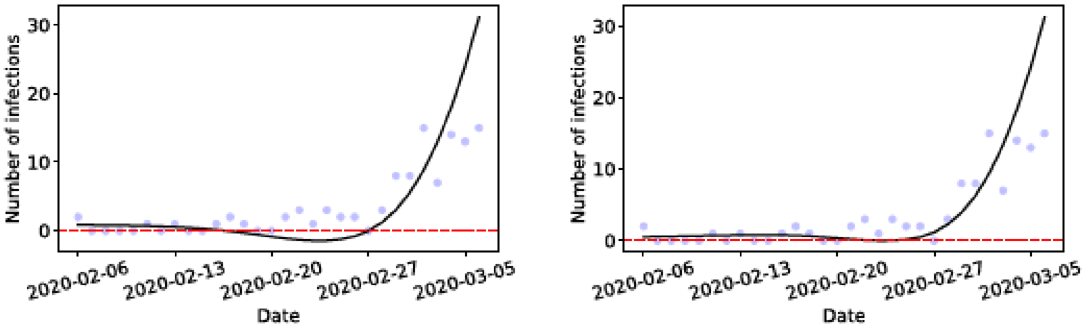


Figure 1: Smoothing and forward prediction curves obtained for the number of daily COVID-19 cases in Aragón, together with the observed values scattered as blue dots. The solid red vertical line depicts the day in which prediction begins.

Figure 3 shows the contour plots of the fitted surfaces using all the data at hand and setting

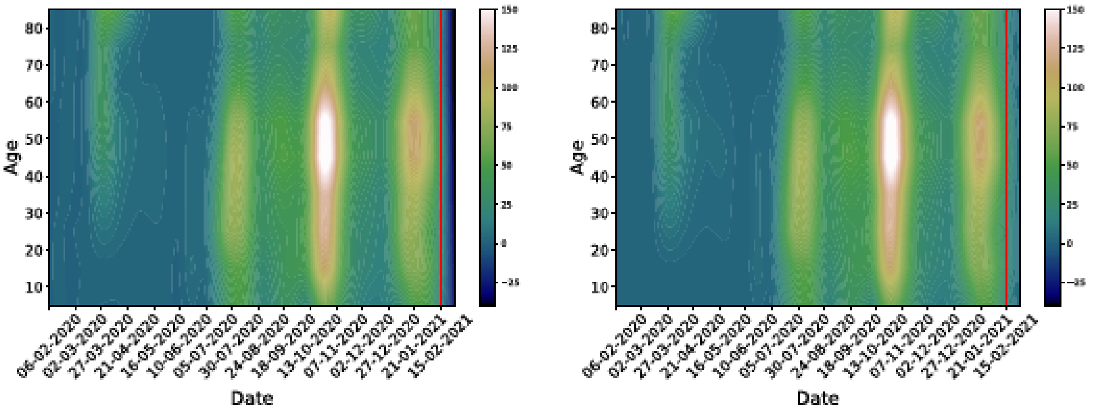


(a) Unconstrained fitted curve.

(b) Non-negative fitted curve.

Figure 2: Zoom in of Figure 1 from February 6th to March 5th 2020 to show the unconstrained and non-negative constrained fitted curves. The red dashed line corresponds to $Y = 0$.

the forecasting horizon to 14 days ahead. When the smoothing is carried out without enforcing any requirements, a clear problem arises at the forecasting region: the surface becomes negative, contradicting the non-negative nature of the data. In order to estimate a non-negative surface, we use the proposed methodology to estimate non-negative curves in a finite set of curves contained in the surface which coincide with the position of the knots.



(a) Unconstrained fitted surface.

(b) Non-negative fitted surface.

Figure 3: Smoothing and forward prediction surfaces for the number of daily infected people in Aragón by age group. The solid red vertical line depicts the day when prediction begins.

4. Bibliography

[1] Bertsimas, D., and Popescu, I. (2002). *On the relation between option and stock prices: a convex optimization approach*. Operations Research, 50(2), 358-374.

[2] Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

[3] Currie, I. D., Durban, M., and Eilers, P. H. (2004). *Smoothing and forecasting mortality rates*. Statistical Modelling, 4(4), 279-298.

[4] De Boor, C., and De Boor, C. (1978). *A practical guide to splines* (Vol. 27, p. 325). New York: Springer-Verlag.

[5] Eilers, P. H., and Marx, B. D. (1996). *Flexible smoothing with B-splines and penalties*. Statistical Science, 11(2), 89-121.

A joint model for the effect of body-weight fluctuation on the risk of death from a frequentist perspective.

Andrea Toloba[a,b], Isaac Subirana[c,b], Guadalupe Gómez[a] and Klaus Langohr[a]

[a] *Universitat Politècnica de Catalunya-BarcelonaTech*
[b] *REGICOR Study Group, Hospital del Mar Medical Research Institute (IMIM)*
[c] *CIBER of epidemiology and public health (CIBERESP)*

Abstract

We propose a joint model able to explain the association between body-weight fluctuation and mortality. Weight over time is adjusted with a linear mixed model, and time-to-death is modelled through a proportional hazards model. An appropriate link function is defined to capture body-weight fluctuation. The motivating dataset is the PREDIMED study, an intervention trial conducted in Spain aiming to the prevention of cardiovascular disease.

Keywords: joint models; shared-parameter models; linear mixed models

1. Introduction

Overweight and obesity are associated with higher risk of cardiovascular mortality and all-cause mortality. Weight loss is commonly prescribed as a lifestyle intervention for these patients. However, weight loss is frequently followed by repeated episodes of subsequent regain of the lost weight, resulting in weight fluctuation. Whether such fluctuations in body weight are associated with worse prognosis is controversial [7], leading to questions about the prudence of recommending weight loss in obese patients.

Previous research studies on that association were based on empirical definitions of body-weight fluctuation, such as the difference between successive recorded values or the variance of all intra-individual recorded weights. More recently, studies have attempted to model that association by means of linear mixed models. In 2019, Cologne [1] proposed to fit a linear mixed model for the subject-specific trajectories of body mass index (BMI) and then explore the association between residual variation and mortality. Later, Willis [5] employed a linear mixed model to approximate the subject-specific slope of BMI, and then quantified the association through a Cox proportional hazards models. Notwithstanding, these approaches do not take into account that linear mixed models assume independence between dropout and the non-recorded consequent measurements (missing at random mechanism), which may be misleading for subjects with an event during follow-up.

In this work, we propose employing a shared-parameter model to capture the dependency between the longitudinal and the survival processes. Parameter estimation is achieved under the frequentist paradigm. In particular, model fitting is conducted with function `jointModel()` from the `JM` [2] package. Besides that, we introduce a new definition of body-weight fluctuation, which constitutes the main novelty of this work. This approach takes advantage of the mathematical definition of acceleration, that may be conceived as fluctuation in the context of body weight. To take into account the history of weight changes that otherwise would be ignored, the fluctuation is averaged over time. For an overview of the joint modelling of longitudinal and time-to-event data, we recommend the work of Rizopoulos [3].

2. Motivating data

The dataset includes information from 6201 people with a mean age of 67 years, from which 2706 (43.64%) are males and 3495 (56.36%) are females. They were followed over a median time of 5 years (IQR=[2.99; 5.85]). During the follow-up, 188 (3.03%) males and 112 (1.81%) females died. Risk factors considered from baseline were age, diabetes, obesity, hypercholesterolemia, hypertension, and smoking status. Body weight was measured annually up to a total of 8 times. The data proceeds from 10 of 11 nodes of the PREDIMED study. For further information, visit `http://www.predimed.es`.

Recent studies have shown that biological sex influences almost all aspects of organism behaviour [6]. Therefore, body-weight fluctuation is expected to behave differently according to sex. With that in mind, we decided to perform separated analysis for males and females.

3. The joint model

Following the notation by Rizopoulos [3], let $m_i(t)$, $i = 1, \dots, n$, denote the true body weight of the i-th participant at time t. It is assumed that the recorded weights are affected by biological variation and measurement error, so the random variable $y_i(t)$ is introduced to denote the hypothetical observed weight at time t. The joint model is composed of a survival submodel and a longitudinal submodel, which are connected through a shared vector of random effects. The longitudinal submodel aims to estimate the subject-specific true body weight from the gathered covariates, denoted by the vectors $x_i(t)$ and $z_i(t)$, and is formulated as a linear mixed model:

$$\begin{cases} y_i(t) & = m_i(t) + \varepsilon_i(t) = x_i'(t)\beta + z_i'(t)b_i + \varepsilon_i(t) \\ b_i & \sim \mathcal{N}(0,\ D) \\ \varepsilon_i(t) & \sim \mathcal{N}(0,\ \sigma^2) \end{cases}$$

The vector of parameters β contains the fixed effects and is shared between all participants, whereas the random variable b_i contains the random effects and follows a multivariate normal distribution with mean 0 and covariance matrix D. Unlike in the general linear mixed model, the residual terms $\varepsilon_i(t)$ are assumed independent and homoscedastic.

Additionally, let T_i^* be the true time of death for the i-th participant. Given the baseline covariates w_i of the individual, we consider the hazard function $\lambda_i(t \mid \mathcal{M}_i(t), w_i)$ of T_i^*, where $\mathcal{M}_i(t) = \{m_i(s),\ 0 \leq s < t\}$ is the subject-specific historical weight up to time t. Hence, the survival submodel is formulated as a proportional hazards model,

$$\lambda_i(t \mid \mathcal{M}_i(t),\ w_i) = \lambda_0(t) \cdot \exp\Big(\gamma' w_i + \alpha_1 m_i(t) + \alpha_2 f\big(m_i(t)\big)\Big),\ \ t > 0,$$

where $\gamma, \alpha_1, \alpha_2$ are the regression parameters and $f(\cdot)$ is the link function. We have settled a piecewise-constant specification for the baseline hazard function $\lambda_0(t)$.

To complete the formulation of the joint model, the link function $f(\cdot)$ must be defined to represent the body-weight fluctuation registered for each participant. Our proposal is to analytically extract the second derivative from the body-weight profile, and then average it, computing the integral over the follow-up time. That is,

$$f(m_i(t)) = \frac{\int_0^t \frac{d^2}{dt^2} m_i(t)}{t}.$$

In essence, it is interpreted as the average fluctuation registered up to time t, according to the longitudinal submodel. For this reason, the adequacy of $f(\cdot)$ to capture body-weight fluctuation relies heavily upon the goodness of fit of the longitudinal submodel. In particular, considering a complex structure for time is essential to reach a good fitting.

3.1. Variable selection

The linear mixed models have been built following the guidelines of Verbeke and Molenberghs [4]. As a result, the longitudinal submodel is (1) for males and (2) for females, where `age` stands for baseline age, `hypchol` for hypercholesterolemia, and `hypten` for hypertension. Analysis of deviance tables were conducted to select the covariates for the proportional hazards model, which can be consulted in Table 1. Residual plots have been checked for model diagnosis.

$$\begin{aligned} m_i(t) = {} & \beta_1 + \beta_2\,\texttt{(age-67)} + \beta_3\,\texttt{hypchol} + \beta_4\,\texttt{hypten} + b_{i1} + \\ & + t \cdot \big(\beta_5 + \beta_7\,\texttt{(age-67)} + \beta_8\,\texttt{hypchol} + b_{i2}\big) + \\ & + (\beta_6 + b_{i3})\,t^2 \qquad (1) \\ m_i(t) = {} & \beta_1 + \beta_2\,\texttt{(age-67)} + \beta_3\,\texttt{hypchol} + \beta_4\,\texttt{hypten} + b_{i1} + \\ & + t \cdot \big(\beta_5 + \beta_9\,\texttt{(age-67)} + b_{i2}\big) + (\beta_6 + b_{i3})\,t^2 + \\ & + (\beta_7 + b_{i4})\,t^3 + \beta_8\,t^4 \qquad (2) \end{aligned}$$

3.2. Results and discussion

For the longitudinal submodel, the estimated standard deviation of the residuals is $\sigma = 2.23$ for males and $\sigma = 2.05$ for females. Hence, the predicted body weights are quite close to the recorded measurements, though weight changes will result smoothed. Model fitting could be improved by considering time-dependent covariates, such as physical activity or energy consumption.

The hazard ratio associated to body-weight fluctuation corresponds to the last row in Table 1. The lack of effect for males could be due to the little fluctuation the longitudinal submodel has found in the data. Besides that, a sensitivity analysis was conducted for females, revealing that the effect dilutes for females under the age of 70 (HR $= 1.70\,[0.69, 4.19]$). This suggests that the results may be biased by reverse causality. In epidemiology, reverse causality refers to the situation in which the disease is causing or affecting the exposure status and not vice versa. Indeed, patients with declining health status are prone to unstable body weight, which is misleadingly interpreted as body-weight fluctuation in our model.

In conclusion, our approach to analyse body-weight fluctuation constitutes an improvement compared to previous strategies. First, the linear mixed model allows us to interpolate the body-weight profile, accounting for subject-specific effects. Also, its formulation provides a solid basis to formalize body-weight fluctuation. Furthermore, the estimation of the linear mixed model under the joint model framework addresses the dependence of the longitudinal measurements on the survival process. Despite this, efforts should focus on differentiating between intentional and unintentional body-weight fluctuation.

Risk Factor	HR	95% CI
age - 67	1.14	[1.09, 1.18]
hypchol	0.93	[0.66, 1.29]
diabetes	1.38	[1.02, 1.85]
smoker	1.65	[1.22, 2.24]
age-67 : hypchol	0.95	[0.90, 0.99]
body weight kg	1.01	[1.00, 1.03]
body-weight fluc.	1.16	[0.45, 3.00]

Risk Factor	HR	95% CI
age - 67	1.11	[1.07, 1.15]
diabetes	1.80	[1.23, 2.62]
body weight kg	1.01	[0.99, 1.04]
body-weight fluc.	2.04	[1.28, 3.27]

Table 1: On the left (resp. right), the results from the survival submodel for males (resp. females).

4. Acknowledgments

I am grateful to the Cardiovascular Epidemiology Department of the Hospital del Mar Medical Research Institute (IMIM) for providing the dataset and research questions, specially to Dr. Montse Fitó and Dr. Helmut Schröder. Also, I want to thank Dr. Klaus Langohr, Dr. Guadalupe Gómez, Dr. Isaac Subirana and Dr. Jaume Marrugat for their support and guidance.

5. Bibliography

[1] J. Cologne, I. Takahashi, B. French, A. Nanri, M. Misumi, A. Sadakane, H. M. Cullings, Y. Araki, and T. Mizoue. Association of weight fluctuation with mortality in japanese adults. *JAMA Network Open*, 2(3):e190731, 2019.

[2] D. Rizopoulos. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33, 2010.

[3] D. Rizopoulos. *Joint models for longitudinal and time-to-event data: with applications in R*. Biostatistics series. CRC Press, 2012.

[4] G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer, 2000.

[5] E. A. Willis, W.-Y. Huang, P. F. Saint-Maurice, M. F. Leitzmann, E. A. Salerno, C. E. Matthews, and S. I. Berndt. Increased frequency of intentional weight loss associated with reduced mortality: a prospective cohort analysis. *BMC Medicine*, 18(1):248, 2020.

[6] S. J. Winham and M. M. Mielke. What about sex? *Nature Metabolism*, 2021.

[7] H. Zou, P. Yin, L. Liu, W. Liu, Z. Zhang, Y. Yang, W. Li, Q. Zong, and X. Yu. Body-weight fluctuation was associated with increased risk for cardiovascular disease, all-cause and cardiovascular mortality: A systematic review and meta-analysis. *Frontiers in Endocrinology*, 10:728, 2019.

Fitting double hierarchical generalized linear models with INLA

Mabel Morales Otero[1], *Virgilio Gómez Rubio*[2], *Vicente Núñez Antón*[3]

[1]mabel.morales@ehu.eus, Universidad del País Vasco (UPV/EHU)
[2]Virgilio.Gomez@uclm.es, Universidad de Castilla-La Mancha (UCLM)
[3]vicente.nunezanton@ehu.eus, Universidad del País Vasco (UPV/EHU)

Double Hierarchical Generalized Linear Models offer great flexibility when modeling highly structured data. We present an approach to fitting these models using a combination of INLA and Importance Sampling, which offers an alternative to other existing estimation methods such as MCMC. We provide specific details on the implementation of the proposed method by performing simulation studies and by applying it to real data examples.

Keywords: Bayesian statistics, DHGLM, INLA, Importance Sampling

1. Introduction

Double Hierarchical Generalized Linear Models (DHGLM) [1] are extensions of Generalized Linear Models (GLM) which allow to model both the mean and the dispersion or scale parameters of the response variable distribution. Moreover, random effects can be included in the regression structures and, in addition, the precisions or variances of these terms can also be modeled. As they are highly structured models, these models can be rather complicated to fit. One option could be to use Markov Chain Monte Carlo (MCMC) methods for their estimation, but it could require a long computation time. An alternative to MCMC is usually given by the Integrated Nested Laplace Approximation (INLA), but it is not possible to fit these models directly with this approach. Therefore, and following Berild et al. (2021) [2], we propose an innovative modelling approach that includes a combination of INLA and a sampling method such as the Adaptive Multiple Importance Sampling (AMIS).

We carried out three different simulation studies to better illustrate model fitting of hierarchical models with different structures using the JAGS software, and compare the results obtained with the MCMC approach. We have also applied this methodology to two real data examples: the study of infant mortality rates in Colombia and to modeling the effect of sleep deprivation in the reaction time on a number of subjects. However, for brevity of exposition, here we only include the results for two of the simulations and one of the real data examples.

2. Methods

DHGLM are specified given a set of two random effects $(\mathbf{u}^{(\boldsymbol{\mu})}, \mathbf{u}^{(\boldsymbol{\phi})})$, so that the conditional mean and variance of the response variables Y_i are $\mathrm{E}[\mathrm{Y_i}|\mathbf{u}^{(\boldsymbol{\mu})}, \mathbf{u}^{(\boldsymbol{\phi})}] = \mu_\mathrm{i}$ and $\mathrm{Var}[\mathrm{Y_i}|\mathbf{u}^{(\boldsymbol{\mu})}, \mathbf{u}^{(\boldsymbol{\phi})}] = \phi_\mathrm{i}\mathrm{V}(\mu_\mathrm{i})$, respectively, for $i = 1, \ldots, n$. The random effects depend on the variance (or the precision) parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$, i.e., $(\mathbf{u}^{(\boldsymbol{\mu})}(\boldsymbol{\lambda}), \mathbf{u}^{(\boldsymbol{\phi})}(\boldsymbol{\alpha}))$. Here, regression models for the mean, for the dispersion parameters and for the parameters of the random effects are specified as:

$$\begin{array}{rcl} g^{(\boldsymbol{\mu})}(\mu_i) & = & \mathbf{X}_i^{\top(\boldsymbol{\mu})}\boldsymbol{\beta}^{(\boldsymbol{\mu})} + \mathbf{Z}_i^{\top(\boldsymbol{\mu})}u_i^{(\boldsymbol{\mu})} \\ g^{(\boldsymbol{\lambda})}(\lambda_i) & = & \mathbf{X}_i^{\top(\boldsymbol{\lambda})}\boldsymbol{\beta}^{(\boldsymbol{\lambda})} \\ g^{(\boldsymbol{\phi})}(\phi_i) & = & \mathbf{X}_i^{\top(\boldsymbol{\phi})}\boldsymbol{\beta}^{(\boldsymbol{\phi})} + \mathbf{Z}_i^{\top(\boldsymbol{\phi})}u_i^{(\boldsymbol{\phi})} \\ g^{(\boldsymbol{\alpha})}(\alpha_i) & = & \mathbf{X}_i^{\top(\boldsymbol{\alpha})}\boldsymbol{\beta}^{(\boldsymbol{\alpha})}, \end{array} \tag{1}$$

where $\mathbf{X}_i^{\top(\cdot)}$ is the i-th row of the design matrix $\mathbf{X}^{(\cdot)}$ for $\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\phi}$ and $\boldsymbol{\alpha}$, $\mathbf{Z}_i^{\top(\cdot)}$ is the i-th row of the design matrix $\mathbf{Z}^{(\cdot)}$ for $\boldsymbol{\mu}, \boldsymbol{\phi}$, and $\boldsymbol{\beta}^{(\cdot)}$ is a vector of unknown coefficients to be estimated for $\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\phi}$ and $\boldsymbol{\alpha}$, respectively.

INLA provides fast approximate Bayesian inference for latent Gaussian Markov random field (GMRF) models, and could be used as an alternative to MCMC methods to fit DHGLM's. However, they do not belong to the class of models that INLA can fit due to their particular structure that includes different hierarchies on the mean and scale parameters. Nevertheless, DHGLM can be expressed as conditional latent GMRF models after conditioning on some model parameters. This idea of fitting conditional models with INLA has been exploited by several authors (see, e.g., [3]) to increase the number of models that can be fitted with INLA. The methodology is based on splitting the vector of hyperameters $\boldsymbol{\theta}$ into two subsets, $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_{-c}$, so that the model, conditional on $\boldsymbol{\theta}_c$, can be fitted with INLA.

In our proposed Importance Sampling (IS) algorithm, samples of $\boldsymbol{\theta}_c$, (i.e., $\{\boldsymbol{\theta}_c^{(m)}\}_{m=1}^{M}$) are obtained using an importance or proposed distribution $s(\cdot)$. For each value $\boldsymbol{\theta}_c^{(m)}$, a conditional model is fitted with INLA and the integration weights w_m, which account for the difference between the proposal and the target distribution, are computed. These samples and weights are used to approximate the target distribution, which is the posterior marginal, using a weighted average. A more detailed description of the method is provided in [2]. AMIS has also been described to provide a more robust sampling method that updates the importance distribution $s(\cdot)$, for which a multivariate Gaussian and a multivariate t distributions have been proposed [2].

3. Simulation Studies and Application

The first simulation study corresponds to a Poisson log-linear model with fixed and random effects, so that the precision of the random effects is modeled by using a linear term with covariates. More specifically, the model is given by:

$$\begin{array}{rcl} Y_i & \sim & \mathrm{Poi}(\mu_{\mathrm{i}}),\ \mathrm{i} = 1, \ldots, \mathrm{n} \\ \log(\mu_i) & = & \beta_0 + \beta_1 x_i + u_i,\ \text{with } u_i \sim N(0, \tau_i) \\ \log(\tau_i) & = & \gamma_0 + \gamma_1 z_i \end{array} \tag{2}$$

For the simulated data, we have used $n = 1000$, $\beta_0 = 1$, $\beta_1 = 0.25$, $\gamma_0 = 0$ and $\gamma_1 = 0.5$. Covariates were simulated so that: $x_i \sim U(0, 1)$ and $z_i \sim N(0, 1)$.

We have also performed a simulation study for a negative binomial model, with different scale parameters for each observation:

$$\begin{array}{rcl} Y_i & \sim & \text{NB}(p_i, k_i),\ i = 1, \ldots, n,\ \text{with}\, p_i = \frac{k_i}{k_i + \mu_i} \\ \log(\mu_i) & = & \beta_0 + \beta_1 x_i \\ \log(k_i) & = & \gamma_0 + \gamma_1 z_i \end{array} \tag{3}$$

We simulated $n = 500$ observations, the values for the parameters were $\beta_0 = 1$, $\beta_1 = 0.25$, $\gamma_0 = 0$ and $\gamma_1 = 5$, and the covariates were simulated so that: $x_i \sim U(10, 20)$ and $z_i \sim U(0, 20)$.

These models can be considered as DHGLM, as the mean and dispersion parameters are being modeled simultaneously. In addition, it could be useful to highlight that models like these are often used to model overdispersed count data [4]. They can be expressed as latent GMRF's by conditioning on $\boldsymbol{\theta}_c = \boldsymbol{\gamma} = (\gamma_0, \gamma_1)$, resulting in a Poisson model having random effects with different precisions, and, in the case of a negative binomial model, with different sizes, which we will fit by using AMIS with INLA. Values of $\boldsymbol{\gamma}$ will be obtained by simulation, and estimates of the posterior distribution by using importance weights. The posterior marginals for β_0 and β_1 will be obtained by respectively weighting their conditional marginals. In both cases, the initial sampling distribution was assumed as a multivariate Gaussian with zero mean and a diagonal variance matrix with entries equal to 5.

Furthermore, we have analyzed data about the effect of sleep deprivation on reaction time for a number of subjects, which can be found in the `R` package `lme4`, and it includes observations for the most sleep-deprived group for the first 10 days of the study. We have fitted a DHGLM with random slopes per subject and different within subject precisions, so that:

$$\begin{array}{rcl} Y_{ij} & \sim & N(\mu_{ij}, \tau_i);\ i = 1, \ldots, p;\ j = 1, \ldots, n_i \\ \mu_{ij} & = & \beta_0 + \beta_i \text{day}_{ij},\ \text{with}\, \beta_i \sim N(0, \tau_\beta) \\ \log(\tau_i) & = & \gamma + u_i,\ \text{with}\, u_i \sim N(0, \tau_u) \end{array} \tag{4}$$

Here, $p = 18$ is the number of subjects and $n_i = 10$, $i = 1, \ldots, p$, given that all subjects have the same number of measurements in the dataset. Covariate day_{ij} is the number of the days that have gone by since the beginning of the sleep deprivation experiment. In order to fit this model by using AMIS with INLA, we condition on $\boldsymbol{\theta}_c = \boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)$, which will split the main model into two independent submodels with response variables $\mathbf{y}$ and $\log(\boldsymbol{\tau})$. These two models can be independently fitted, and the resulting log-marginal likelihood will be the sum of the corresponding values from the two models, which can be then used to compute the weights in the AMIS algorithm. For the importance distribution, we propose to use a multivariate Gaussian with parameters informed from the data. That is, the mean is $\left[\log(1/S_1^2), \ldots, \log(1/S_p^2)\right]$ and the variance is a diagonal matrix with entries
$0.05 \times \left[\log(1/S_1^2), \ldots, \log(1/S_p^2)\right]$, where S_i^2 is the sample variance for the measurements on the i-th individual. These initial values provided better results than those using a vague starting distribution. Tables 1, 2 and 3 summarize the estimates obtained by using both methods for each of the models considered here. As can be seen, for all the models considered here, the results from the MCMC estimations are very close to those obtained with AMIS-INLA.

		AMIS		MCMC	
Parameter	True value	Mean	St. dev.	Mean	St. dev.
β_0	1	1.0531	0.0736	1.0509	0.0745
β_1	0.25	0.2302	0.1254	0.2286	0.1270
γ_0	0	-0.0210	0.0655	-0.0455	0.0651
γ_1	0.5	0.4830	0.0622	0.4787	0.0628

Table 1: Summary of the estimates of the Poisson log-linear model used in the simulation study.

		AMIS		MCMC	
Parameter	True value	Mean	St. dev.	Mean	St. dev.
β_0	1	0.9875	0.0541	0.9941	0.0513
β_1	0.25	0.2506	0.0033	0.2502	0.0031
γ_0	0	-0.0879	0.0926	-0.0850	0.0903
γ_1	5	4.8594	0.1861	4.8588	0.1813

Table 2: Summary of the estimates of the negative binomial model used in the simulation study.

	AMIS		MCMC	
Parameter	Mean	St. dev.	Mean	St. dev.
β_0	0.2606	0.0034	0.2590	0.0042
τ_β	8240.2229	2742.75	8005.402	2712.491
γ	7.3170	0.2003	7.2666	0.2150
τ_u	2.1222	0.9348	2.6477	2.5713

Table 3: Summary of the estimates of the Gaussian model fitted to the sleep study data.

4. Discussion

We have illustrated model fitting of DHGLM by conducting two different simulation studies and by performing the analysis of a real data example, including also suggestions on how to choose the initial values for the algorithm. In all cases, AMIS-INLA provided good estimates of the model effects and hyperparameters, which were very close to those obtained with MCMC methods. Hence, we can conclude that our proposed AMIS-INLA approach provides an efficient way of fitting DHGLM, representing a valid alternative to the MCMC estimation methods.

5. Bibliography

[1] Lee, Y. and Nelder, J.A. (2006). *Double hierarchical generalized linear models (with discussion).* Applied Statistics, 55(2), 139–185.

[2] Berild, M., Martino, S., Gómez-Rubio, V. and Rue, H. (2021). *Importance sampling with the*

integrated nested Laplace approximation. arXiv:2103.02721 [stat.CO].

[3] Gómez-Rubio, V. and Rue, H. (2018). *Markov chain Monte Carlo with the integrated nested Laplace approximation*. Statistics and Computing, 28(5), 1033–1051.

[4] Quintero-Sarmiento, A., Cepeda-Cuervo, E. and Núñez-Antón, V. (2012). *Estimating infant mortality in Colombia: some overdispersion modelling approaches*. Journal of Applied Statistics, 39(5), 1011–1036.

Interactive modelling and prediction of patient evolution via multistate models

Leire Garmendia Bergés[1]*; Jordi Cortés Martínez*[2]*; Guadalupe Gómez Melis*[3]

[1]leire.garmendia@upc.edu, Universitat Politècnica de Catalunya
[2]jordi.cortes-martinez@upc.edu, Universitat Politècnica de Catalunya
[3]lupe.gomez@upc.edu, Universitat Politècnica de Catalunya

Abstract

A friendly interactive web app has been built to allow to fit a multistate model from specific data and to predict the clinical evolution for a given patient. The app, amenable to other diseases with different stages of severity, allows to visualize the disease process easing the communication among members of different fields and turning it into a very useful tool for the clinical and logistical management of hospitalized patients.

Keywords: Multistate model, shiny app, COVID-19

1. Introduction

Modelling the course of a disease regarding severe events and identifying prognostic factors is of great clinical relevance. Besides death, other intermediate events indicative of disease progression are relevant for clinical management. Multistate models (MSM) can be used to model the movement of patients among several states [1]. Specifically, they are useful to analyse a disease with an increasing degree of severity, that may precede death [2]. MSM interests include the estimation of progression rates, evaluating the associations of risk factors, survival rates or prediction. MSM offer several advantages with respect to other methods in survival analysis. They are more flexible than Cox models with time-dependent covariates since they allow different baseline hazards for each transition. In addition, they are more intuitive than landmark models because all prediction timepoints are included in a single model and since different covariates can be specified in each transition. But, the main advantage compared to both of these methods is that the model provides estimates of probabilities of being in both intermediate and absorbing states at any point in time. Moreover, the visual representation of a MSM generally makes it more interpretable for clinical researchers. Motivated by the strengths of this type of model, this work presents a web app with two main goals: 1) to provide the user with the ability to fit a MSM in a friendly way; 2) to predict the clinical evolution for a given patient based on the previous model. The app will be illustrated to predict high-risk critically ill cases among COVID-19 hospitalized patients.

2. Methodology

A MSM is a generalization of classic survival analysis that makes possible to describe complex clinical changing processes over time. It is defined by a series of states — which could represent different stages of the disease — and the transitions that connect them — which trace the evolution of a patient during a follow-up period. MSM allow complex settings such as the one to describe the evolution of a patient that is admitted into the hospital due to COVID-19 as we encounter in the DIVINE (*DynamIc eValuation of COVID-19 cliNical statEs and their prognostic factors to improve the intra-hospital patient management*) project. Figure 1 shows the schema adopted for these patients.

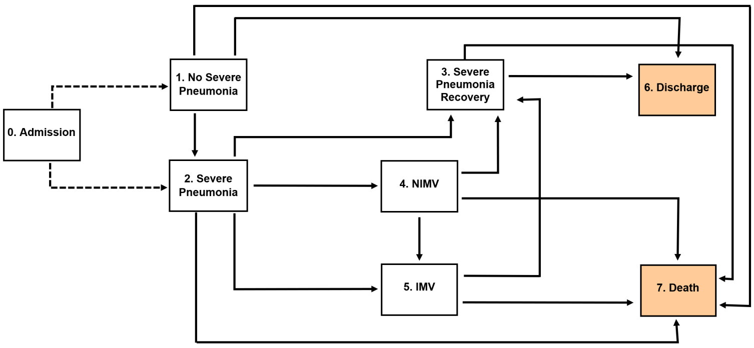


Figure 1: MSM in DIVINE project

A multistate process is a model for a continuous-time stochastic process $\{X_t; t \leq \tau\}$, which at any time point occupies one of a set of discrete states $\mathcal{S}$ ($\mathcal{S} = 1, .., k$). For all $i, j \in \mathcal{S}$ the probability of transition to state j at time t, knowing that the patient was in state i at time s (for $s < t$) is defined as:

$$P_{ij}(s, t; \mathcal{F}_{s^-}) = P(X_t = j | X_s = i; \mathcal{F}_{s^-}) \tag{1}$$

where $\mathcal{F}_{t^-}$ represents the history of the patient up to time t. For all $i, j \in \mathcal{S}$ the transition intensity (or the instantaneous hazard) between two states i and j is defined as:

$$\lambda_{ij}(t; \mathcal{F}_{t^-}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P_{ij}(t, t + \Delta t; \mathcal{F}_{t^-}) \tag{2}$$

Under the Markov assumption, i.e., that the future only depends on the present, not on the past, and the one we will use for this web, expressions (1) and (2) are simpler:

$$\begin{aligned} P_{ij}(s,t) &= P(X_t = j | X_s = i), \ \ \forall i, j \in \mathcal{S} \\ \lambda_{ij}(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P_{ij}(t, t + \Delta t), \ \ \forall i, j \in \mathcal{S} \end{aligned}$$

Besides the nonparametric estimation of $P_{ij}(s, t)$ and $\lambda_{ij}(t)$ interest relies on the effect that covariates could have of them. We will focus on the semi-parametric choice via Cox (proportional hazards) models allowing to relate the characteristics of an individual, described by some covariates, with the transition intensities as follows:

$$\lambda_{ij}(t; Z) = \lambda_{ij,0}(t) exp\{\beta_{ij}^T Z\} \tag{3}$$

where $\lambda_{ij,0}(t)$ is the baseline hazard function between states i and j, β_{ij} is the vector of regression parameters, and Z is the covariate vector.

3. Web app

A web app created with the *shiny* R package is presented. It has two main features: 1) to allow to fit a MSM from specific data; 2) to predict the clinical evolution for a given patient. To fit the model, the data to be analysed must be upload in a prespecified format. Then, the user have to define the states and transitions as well as the covariates (e.g., age or gender) involved in each transition. From this information, the app returns descriptive graphics to represent the distributions of the selected covariates and the length of stay for each state. To make predictions, the values of selected covariates from a new patient at baseline has to be provided. From these inputs, the app provides some indicators of the patient's evolution such as the probability of 30-day death or the most likely state at a fixed time. Furthermore, visual representations (e.g., the stacked transition probabilities plot) are given to make predictions more understandable. To achieve those goals, the app has different tabs for several functionalities: 1) ***Home***: how the app works along with brief instructions on how to upload a dataset; 2) ***Data***: the user can decide between working with the *example dataset* (from the DIVINE project), which is the default option, or uploading its own data; 3) ***Model***: specification of the MSM by the definition of states, transitions and covariates as well as the analysis type; 4) ***Descriptive analysis***: graphical representations of the distributions of the covariates included in the model as well as the time until/in each state; 5) ***Fitted model***: estimates from the model and their interpretation; 6) ***Output graphics***: instantaneous hazards and transition probabilities over time; 7) ***Prediction***: forecasting for patient's evolution from information on the covariates associated to a new individual; 8) ***Help***: extensive instructions on the app operation.

4. Case study: DIVINE project

A multicohort study of more than 4,000 hospitalized adult COVID-19 patients from 8 Catalan hospitals during the 1st, 2nd, 3rd, and 5th Spanish waves of the pandemic was collected. As the *example data* comes from this project, it is not necessary to upload any file. The app shows the first registers in the *Data* tab.

The model in Figure 1 depicts the most clinically relevant disease severity states to explain the patient's evolution and the meaningful transitions between then based on the agreement of the DIVINE team. In particular, we have the following **States**: *Transient* (non-severe pneumonia, severe pneumonia, non-invasive mechanical ventilation, invasive mechanical ventilation, severe pneumonia recovery) or *Absorbent* (discharge and death); **Transitions**: transitions have to be specified one by one leading to the transition matrix required to fit the model; **Covariates**: age and sex were considered as covariates in this work despite other potential covariates (e.g., oxygen saturation) could be taken into account. In our study, there are about 60% men ranging from 18 to almost 100 years old with mean around 60 years.

Before fitting the model, the app shows descriptive graphs of the selected covariates such as bar charts or histograms, and boxplots representing the stay lengths for each transient state and the time until the absorbing states.

The information regarding the states and the transitions makes up the transition matrix needed to fit the model. The summary of the MSM and a help box with the interpretation of the coefficients are provided. As an example, we comment a couple of preliminary results. The coefficient $\hat{\beta}^M_{1,2} = 0.35$ is related to the sex category *male* (M) for the transition from *non-severe pneumonia* (state 1) to *severe pneumonia* (state 2). The hazard ratio, $HR^M_{1,2} = e^{0.35} = 1.41$; $95\%CI = (1.15, 1.73)$, indicates that being male is associated with a higher risk of moving from *non-severe pneumonia* to *severe pneumonia*. On the other hand, the coefficient $\hat{\beta}^A_{12} = 0.0051$ yields to a hazard ratio for *age* (A), $HR^A_{1,2} = e^{0.0051} = 1.005$; $95\%CI = (0.998, 1.012)$, for the same transition as before. In this case the 95% confidence interval informs that there is no evidence of association between age and an increasing or decreasing risk for a patient to move from *non-severe pneumonia* to *severe pneumonia*.

Some graphs to illustrate the main model indicators (e.g., instantaneous hazard or transition intensities) are given. In our case, we can obtain the instantaneous hazards for the transitions starting in a specific selected transient state. For categorical explanatory variables, the indicators are provided for each level and for continuos covariates, they are dichotomized according to the median. Figure 2 shows how the app allows to distinguish different patterns according to the categories of sex and age over time.

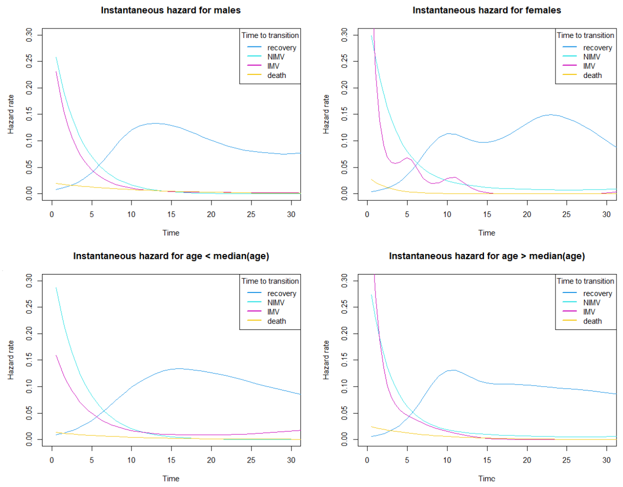


Figure 2: Instantaneous hazards over time in the *Output graphics* tab according to sex (upper panels) and age (bottom panels)

5. Conclusions

Two main strengths over other similar existing apps are: 1) predicts the evolution of a new patient; 2) allows to compare different MSM. Having an interactive app that allows to visualize this process eases communication among members of different fields and represents a very useful tool for the clinical and logistical management of diseases with different stages of severity.

6. Acknowledgments

This research was funded by the Ministerio de Ciencia e Innovación (Spain) [PID2019104830RBI00]; and the Generalitat de Catalunya (Spain) [2017SGR622 and 2020PANDE00148].

7. Bibliography

[1] Cook, R.J.; Lawless, J.F., (2020). *Multistate models for the analysis of life history data* Chapman and Hall.

[2] Mody, A., Lyons, P.G., Vazquez Guillamet, C., et al., (2020). *The Clinical Course of Coronavirus Disease 2019 in a US Hospital System: A Multistate Analysis* American Journal of Epidemiology, 190(4).

Parallel session: Epidemiological studies

Chairperson:

Miguel Angel Martínez, Universitat de Valencia (Spain)

Plasma metabolomics, bone mineral density and fractures in a general population from Spain: The Hortega Study

Marta Galvez-Fernandez[1], Zulema Rodriguez-Hernandez[1], Maria Grau-Perez[2], F.Javier Chaves[3], Daniel Monleon[3], Maria Tellez-Plaza[1] Josep Redon[3], Juan C. Martin-Escudero[3]

[1] National Center of Epidemiology, Health Institute Carlos III
[2]Institute for Biomedical Research, Hospital Clinic of Valencia (INCLIVA)
[3]Hospital Universitario Rio Hortega. University of Valladolid, Spain

Background and objectives: Evidence suggests that the bone remodeling process may be influenced by metabolic factors. We evaluated the cross-sectional association between metabolic profiles with reduced bone mineral density (BMD) and the prospective association of metabolic profiles with incident osteoporosis-related fractures in a representative sample from Spain. Since redox and inflammation status are affected by the metabolomic profile and play an active role in bone resorption, as secondary analyses we assessed the interaction role of redox-related genes.

Methods: In 507 participants older than 50 years from the Hortega Study, we estimated metabolic principal components (mPC) from 54 plasma metabolites, measured with targeted NMR-spectrometry. BMD was calculated in the right calcaneus using Peripheral Instantaneous X-ray Imaging system (PIXI). Redox-related candidate SNPs (N=291) were measured by oligo-ligation assay.

Results: mPC1 (reflecting non-essential and essential amino acids, including branched-chain and bacterial co-metabolism versus fatty acids and VLDL subclasses) was inversely associated with reduced BMD (T-score < -1 SD). mPC2 (reflecting essential aromatic amino acids and bacterial co-metabolism) was positively associated with reduced BMD. The corresponding prospective associations with incident fractures were consistent for these mPCs. mPC4 (reflecting HDL subclasses) was inversely associated with incident osteoporotic fractures. In flexible dose-response analyses mPC3 (reflecting LDL subclasses) was suggestively associated with higher risk of reduced BMD. Carriers of 7 genetic variants showed differential associations between mPCs and reduced BMD, being *ACE* and *NOX1* the most interacting genes.

Conclusions: Our results support the hypothesis that bone remodeling is influenced by metabolic factors, including amino acid and lipids and microbiota co-metabolism. Carriers of redox-related variants may benefit from intensified preventive interventions to prevent bone disease in ageing populations.

Keywords: fractures, bone, metabolomic

Estimation of excess mortality in 2020 in five European countries

Garyfallos Konstantinoudis[1], *Michela Cameletti*[2], *Virgilio Gómez-Rubio*[3], *Inmaculada León Gómez*[4,5], *Monica Pirani*[1], *Gianluca Baio*[6], *Amparo Larrauri*[4,5], *Julien Riou*[7], *Matthias Egger*[7,8], *Paolo Vineis*[1], *Marta Blangiardo*[1]

[1]MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [2]Department of Economics, University of Bergamo, Bergamo, Italy. [3]Departamento de Matemáticas, E.T.S.I.I., Universidad de Castilla-La Mancha, Albacete, Spain. [4]National Centre of Epidemiology (CNE), Institute of Health Carlos III, Madrid, Spain. [5]Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Institute of Health Carlos III, Madrid, Spain. [6]Department of Statistical Sciences, University College London, London, UK. [7]Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland. [8]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK.

We have studied the impact of the COVID-19 pandemic on excess mortality in 2020. All casues of death have been considered in the analysis and observed number of deaths have been compared to the expected deaths under the counterfactual of no-pandemic. Expected counts have been estimated using population and mortality data from period 2014-2019 and Bayesian spatio-temporal models. Average weekly temperature and holidays have been included in the model as covariates to improve model fitting. Our analysis includes 5 European countries and different administrative levels as well as stratification by age group and sex. The results show an increase in mortality in all studied countries, with Comunidad de Madrid, Castilla-La Mancha and Castilla y León with the highest excesses with a ˜30% increase in mortality in 2020. We have developed an App to visualize the different excess mortality estimates and it is available at `http://atlasmortalidad.uclm.es/excess`.

References

Konstantinoudis, G., Cameletti, M., Gómez-Rubio, V. et al. Regional excess mortality during the 2020 COVID-19 pandemic in five European countries. Nature Communications 13, 482 (2022). `https://doi.org/10.1038/s41467-022-28157-3`.

Keywords: Bayesian inference, COVID-19, excess mortality, spatio-temporal models

A proposal for spatial small area estimation in survey studies

Miguel A. Martinez-Beneito[1], *Carlos Vergara-Hernández*[2], *Marc Marí-Dell'Olmo*[3], *Laura Oliveras*[4]

[1]miguel.a.martinez@uv.es, Department of Statistics and Operations Research, University of Valencia

[2]carlos.vergara@uv.es, Department of Nursery, University of Valencia

[3]mmari@aspb.cat, Environmental Quality and Intervention Service, Agència de Salut Pública de Barcelona

[4]lolivera@aspb.cat, Environmental Quality and Intervention Service, Agència de Salut Pública de Barcelona

Spatial statistics for lattice data and survey-based small area estimation have evolved as parallel, but mostly independent, research lines. Although both of them pursue similar goals, deriving reliable estimates over sets of small areas, they have focussed on different statistical approaches. On one hand, spatial statistics have focussed on taking profit of the spatial dependence that data frequently show for deriving enhanced estimates. On the other hand, the small area estimation literature usually focus on using the information on the sampling design of the corresponding survey study so that the corresponding estimates take into account that particular design. Surprisingly, few proposals join both modeling resources (spatial dependence and taking profit of the sampling design), in order to derive spatial survey-based small area estimates.

In this work we propose a model that tries to merge the spatial and survey-based small area approaches within a common Bayesian proposal. Our proposal models, at the individual level, the effect on the outcome of stratification and other additional variables whose totals are known for each unit of study. In addition, spatial random effects are also considered for modeling the spatial variability of the outcome. Finally, the model reweights the individual estimates (with spatial terms) according to the sampling design information yielding survey-based spatial estimates. As shown, this proposal is able to reproduce combined spatial-stratified-ratio estimators from a single model. As an illustration we apply our proposal to the estimation of low- and middle-income countries population in Barcelona neighborhoods and assess the benefits of this approach as compared to more traditional proposals.

Keywords: spatial statistics; small area estimation; bayesian hierarchical models

Recalibration of glucose and insulin determinations to evaluate differential associations of selenium with insulin resistance and β-cell function by age in two study populations

Zulema Rodriguez-Hernandez[1,2*], *Maria Grau-Perez*[3,4*], *Javier Bel-Aguilar*[3], *Josep Redon*[4], *Jose L. Gomez-Ariza*[5], *Tamara Garcia-Barrera*[5], *Belén Callejón-Leblic*[5], *Belén Moreno-Franco, Martin Laclaustra*[6], *Jose Puzo*[6], *Jose A. Casasnovas*[6], *Rosario Ortola*[7], *Fernando Rodriguez-Artalejo*[7], *Esther Garcia-Esquinas*[1,7], *María Tellez-Plaza,*[1,7], *Roberto Pastor-Barriuso*[1]

* Equal author contribution; [1]National Center for Epidemiology, Carlos III Health Institutes, Madrid; [2]Universitat Politècnica de València, Valencia, [3]University of Valencia, Valencia; [4]Biomedical Research Institute INCLIVA, Valencia; [5]University of Huelva, Huelva; [6]IIS Aragon, CIBERCV, Zaragoza University, Zaragoza; [7]Department of Preventive Medicine and Public Health, Universidad Autónoma de Madrid-Idipaz and CIBERESP, Madrid.

Keywords: regression-based recalibration, selenium, insulin resistance.

Introduction: Epidemiological studies identified positive associations between Selenium (Se) and type 2 diabetes. However, the role of Se in insulin resistance and β-cell dysfunction has rarely been investigated. We evaluated whether the association of Se exposure with insulin resistance and β-cell function, measured by HOMA-IR and HOMA-β respectively, in diabetes-free participants is differential by age, in two study populations of middle-age and elderly participants.

Methods: A total of 1163 participants from the Aragon Workers Health Study (AWHS), age range 42-56, and 915 from Seniors-ENRICA-2 (SEN-2) study, age range 64-82, were included in the cross-sectional analysis. To evaluate differential associations by age, independently of laboratory artefacts, we used a regression-based recalibration method using as reference two different age groups of NHANES 2011-2016 population: aged 40-59 years (N=736) and 60-80 years (N=765). All participants were diabetes-free. The recalibration models included age, sex, BMI, HDL cholesterol, and fasting time covariates. We first fitted separate regression models of glucose and log-transformed insulin levels in AWHS and SEN-2 participants, and computed each participant's standardized residuals r_i by dividing model residuals by the estimated residual standard deviation. We then fitted the same models in the two NHANES groups and extracted the model coefficients b_j^{ext} for each covariate and the residual standard deviation s^{ext}. The recalibrated glucose and log-insulin levels were then $y_i^{\text{rec}} = \sum_j b_j^{\text{ext}} x_{ij} + s^{\text{ext}} r_i$, where x_{ij} was their covariate pattern. Finally, HOMA-IR and HOMA-β were calculated using standard formulas.

Results: Median for original HOMA-IR levels were 1.55 (1.08; 2.21) and 2.17 (1.45; 3.11) in AWHS and SEN-2, respectively, and for original HOMA-β were 70.0 (49.8; 100.5) and 135.0 (93.9; 201.1), respectively. For recalibrated HOMA-IR levels were 1.82 (1.21; 2.88) and 2.05 (1.34; 3.09) in AWHS and SEN-2, respectively, and for recalibrated HOMA-β were 74.2 (50.8; 115.3) and 78.8 (54.5; 116.3), respectively. In subsequent association analysis in AWHS, the GMR (95% CI) comparing the 90th and 10th percentiles of Se distribution were 1.07 (0.99, 1.15) and 1.14 (1.06, 1.24) for original HOMA-IR and HOMA-β, respectively, which is consistent with the recalibrated HOMA-IR models. In SEN-2, the corresponding GMR (95% CI) were 1.10 (1.00, 1.21) and 0.93 (0.84, 1.03) for original HOMA-IR and HOMA-β, respectively. Similar results were obtained with recalibrated HOMA-IR.

Conclusions: The positive association of Se with HOMA-IR in both studies suggests that high levels of Se exposure are related with increased insulin resistance. Alternatively, the positive association of Se with HOMA-β in middle aged adults from AWHS but not in older adults from SEN-2, might suggest that increased insulin resistance induces compensatory increased β-cell function in younger ages, being this compensatory capacity decreased with aging. Results from our recalibrated data support that differential associations by age are not due to between-laboratory variation in glucose and insulin measures.

Multiple mediation for uncausally correlated mediators in survival analysis

Arce Domingo-Relloso,[1,2,3] Allan Jerolon,[4] Ana Navas-Acien,[3] Maria Tellez-Plaza,[1] Jose Bermudez[2]

[1] Department of Chronic Diseases Epidemiology, National Center for Epidemiology, Carlos III Health Institute, Madrid, Spain.

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

[3] Department of Environmental Health Sciences, Columbia University Mailman School of Public Health, New York, NY, USA.

[4] Laboratory MAP5, Universite Paris Descartes and CNRS, Paris, France.

Introduction. Mediation analysis aims to quantify to which extent the relationship between two variables happens through intermediate variables. The association between an exposure and a health outcome might be mediated by several correlated biological pathways. However, individual mediated effects cannot be identified conducting individual mediation models when mediators are correlated. We extended the *multimediate* algorithm, which conducts multiple mediation analysis in the setting of uncausally correlated mediators, to survival data. Furthermore, we used real data from the Strong Heart Study to study the potential mediating role of DNA methylation sites, which are generally highly correlated, on the association between smoking and lung cancer.

Methods. The *multimediate* method uses a quasi-Bayesian algorithm to estimate direct, indirect and total effects based on the simulation of counterfactual distributions. We extended this algorithm, which is implemented in the multimediate R package, to a survival context using additive hazards models. The SHS is a prospective cohort study of American Indian adults. DNA methylation was measured in 2,351 participants at the baseline visit (1989-1991) using the Illumina MethylationEPIC BeadChip (850K). The preprocessing resulted in data from 2,235 individuals and 788,368 methylation sites. Lung cancer incidence was assessed through 2017. Smoking status and cumulative smoking dose were self-reported.

Results. There were 97 lung cancer cases. The individual mediation models identified 20 methylation sites as having significant mediated effects on the association between smoking and lung cancer. Many of them were annotated to well-known smoking-related genes such as *AHRR, F2RL3* and *PRSS23.* The multiple mediation model identified a joint mediated effect of methylation sites of 292 lung cancer cases attributable to current vs. never smoking through DNA methylation changes per 100,000 person-years (79.1 % mediated percentage).

Conclusions. Using a semiparametric additive hazards model, we extended a multiple mediation approach for uncausally correlated mediators to time-to-event outcomes. Our results suggest that 79 % of the effect of smoking in lung cancer might be mediated by changes in DNA methylation in well-known smoking-related genes. Our study provides evidence in favor of DNA methylation as a potential biomarker for early diagnosis of lung cancer, and as a potentially impactful target for future epigenetic treatments.

Parallel session: Medical studies 1

Chairperson:

Xabier Barber, Universidad Miguel Hernández (Spain)

A new general and multivariable approach to categorize predictor variables. Optimal categorization of physical activity in COPD patients

Irantzu Barrio[1], *Javier Roca-Pardiñas*[2], *Cristobal Esteban*[3], *Maria Durban*[4]

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

[2]roca@uvigo.es, Department of Statistics and Operations Research, University of Vigo

[3]cristobal.estebangonzalez@osakidetza.eus, Respiratory Department, Hospital Galdakao-Usansolo, Galdakao, Bizkaia, Spain

[4]mdurban@est-econ.uc3m.es, Department of Statistics and Econometrics, Universidad Carlos III de Madrid

The use of discretized variables in the development of prediction models is a common practice, partly because the decision-making process is more natural when based on rules created from segmented models. Although this practice is perhaps most common in medicine, it is extensible to any area of knowledge in which a prediction model supports the decision-making process. Therefore, providing researchers with a useful and valid categorization method may be a relevant issue when developing predictive models.

A methodology to select the optimal cut-off points to categorize a continuous covariate has been previously proposed in the context of logistic and Cox regression models, which was based on the maximization of the discrimination ability of the model measured by the c-index.

In this work, we propose a new general methodology, which can be applied to categorize a predictor variable in any regression model where the response variable belongs to the exponential family distribution. This new methodology can be applied in any multivariate contexts, moreover, allows to estimate cut-off points deferentially for the different levels of a factor variable with which the variable to be categorized has a significant interaction. Furthermore, a computationally very efficient method is proposed to obtain the optimal number of categories. Several simulation studies have been conducted in which the efficiency of the method with respect to both the location and the number of estimated cut-off points is shown.

Finally, we applied this proposal to a real data set of patients with stable chronic obstructive pulmonary disease (COPD). Accurate assessment of measured physical capacity such as hand, quadriceps or shoulder strength and the intensity of physical activity in daily life is considered very important due to the close relationship between this parameters and patients health and mortality. Therefore, we have applied the proposed methodology in order to obtain optimal categorization proposals for continuous predictor variables related to patient's physical activity considering mortality and number of hospitalizations as outcome variables.

Keywords: optimal categorization, COPD, physical activity.

Could mortality data improve short-time cancer incidence predictions?

Garazi Retegui[1,2,3], *Jaione Etxeberria*[2,3], *María Dolores Ugarte*[2,3], *Andrea Riebler*[4]

[1]garazi.retegui@unavarra.es

[2] Department of Statistics, Computer Science and Mathematics, Public University of Navarre (UPNA), Arrosadia Campus, 31006, Pamplona, Spain

[3]Institute for Advanced Materials and Mathematics (INAMAT2), Public University of Navarre (UPNA), Arrosadia Campus, 31006, Pamplona, Spain

[4]Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Alfred Getz' vei 1, 7034, Trondheim, Norway

Cancer incidence figures play an important role in the allocation of resources aimed at preventing and controlling the disease. They are routinely recorded by national or regional population-based cancer registries. However, these figures are usually available with a delay and therefore, methods that provide short-term predictions are very useful. On top of that, an additional drawback is encountered. In large countries, regional cancer registries are responsible for collecting and identifying all cancer cases occurring in a certain domain (state or province for example). Regional registries are usually not established in the same year and therefore, cancer incidence data series between different regions of a country are not harmonised over time. This lack of information (mainly at the beginning of the data series), makes it difficult to use univariate spatio-temporal models. One possibility to solve this problem is to use cancer mortality data as an additional source of information because cancer incidence and cancer mortality are highly related.

The aim of this work is twofold: on the one hand, to predict cancer incidence in the short-term and, on the other hand, to complete the data series of the cancer registries that are incorporated at different times throughout the studied period. Here we use different multivariate spatio-temporal models. The performance of these multivariate models will be analysed using lung and prostate cancer incidence data during the period 2001-2015 reported by the 16 regional cancer registries of Germany and provided by the German Centre for Cancer Registry Data, ZfKD.

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

A two stage joint modelling of longitudinal discrete bounded outcomes and survival analysis: an application to chronic obstructive pulmonary disease patients

Galán García-Arcicollar, Cristina[1], *Najera-Zuloaga, Josu*[2], *Arostegui, Inmaculada*[3], *Cristobal Esteban*[4], *Lee, Dae-Jin*[5]

[1]cgalan@bcamath.org, Basque Center for Applied Mathematics, Bilbao, Spain
[2]jnajera@deusto.es, Department of Mechanics, Design and Industrial Organization, Universidad de Deusto, Bizkaia, Spain
[4]inmaculada.arostegui@ehu.eus, Department of Mathematics, University of the Basque Country UPV/EHU and Basque Center for Applied Mathematics
[4]cristobal.estebangonzalez@osakidetza.eus, Pneumology Service, Galdakao Hospital, Bizkaia, Spain
[5]dlee@bcamath.org, Basque Center for Applied Mathematics, Bilbao, Spain

In recent years, there has been an increased focus on placing patients at the center of health care research and evaluating clinical care. For instance, patient-reported outcomes (PROs) are helpful tools that provide reports about patient's health status considering their health, quality of life, or functional status associated with the health care or treatment they received. PROs come directly from the patient, without any intervention from a clinician, and are now widely utilized for routine monitoring and assessment of care outcomes in adult patients. Most clinical studies involve the follow-up of patients where longitudinal information is collected. PROs can be considered as discrete and bounded random variables, and hence a Beta-Binomial distribution can be used to estimate regression models with longitudinal information and evaluate the temporal evolution of patients' health-related quality of life. When survival (or time-to-event) analysis is also considered, the statistical literature uses the term "joint modelling" to refer to methods for simultaneously analyzing longitudinal measurement outcomes and time-to-event outcomes.

In this work, we aim to examine the effect of PROs on the patient's risk, so we propose a joint model where the longitudinal component consists of a model for the PROs that might be highly associated with overall survival. Therefore, the first step consists of fitting a longitudinal Beta-binomial mixed-effects model and then in a second step, included the estimated linear predictor in a Cox proportional hazard regression model. The approach is applied to a study of 543 chronic obstructive pulmonary disease (COPD) patients from Galdakao hospital in Spain that includes the specific PRO survey for COPD patients, the St. George Respiratory Questionnaire (SGRQ), consisting of three scales Activity, Symptoms and Impacts.

Keywords: Joint modelling, Beta-Binomial longitudinal model, Survival analysis.

AMS: 62N02 - Estimation in survival analysis and censored data

1

Statistical Learning Models in Classifying the Type of Meningitis

Irene García Mosquera[1],

[1]irene.garcia@uib.es, Department of Mathematical Sciences and Informatics, University of the Balearic Islands (UIB), and Balearic Health Research Institute (IdISBa),

2

The differential diagnosis between bacterial and viral meningitis is a very important problem: On one hand, failure to deliver proper antibiotic therapy in bacterial meningitis can lead to severe, permanent sequelae, and on the other hand, unnecessary antibiotic or overtreatment of viral meningitis cases can lead to antimicrobial resistance, increased health care services cost, changes in human microbiome, and high levels of stress to the suffering patients.

I present the results obtained when applying three statistical learning (SL) algorithms (multiple logistic regression (MLR), random forest (RF), and naïve-Bayes (NB)) for the differential diagnosis of patients with meningitis. Cerebrospinal fluid (CSF) neutrophils, CSF lymphocytes, neutrophil-to-lymphocyte ratio (NLR), blood albumin, blood C-reactive protein (CRP), glucose, blood soluble urokinase-type plasminogen activator receptor (suPAR), and CSF lymphocytes-to-blood CRP ratio (LCR) were used as covariates for the SL algorithms. The performance of the SL algorithms was evaluated through a cross-validation procedure, and optimal predictions of the type of meningitis were above 95% for viral and 78% for bacterial meningitis.

Keywords: Statistical Learning, Differential Meningitis.

Confounding by indication: the role of clinical stability on antimicrobial de-escalation in community-acquired pneumonia

Esther García-Lerma[1] Natalia Pallarés[1] Gabriela Abelenda-Alonso[2], Carlota Gudiol[2], Cristian Tebé[1], Jordi Carratalà[2]

1 egarcia@idibell.cat, npallares@idibell.cat, ctebe@idibell.cat, Biostatistics Unit, Bellvitge Biomedical Research Institute (IDIBELL)
2 gabi.abelenda.alonso@gmail.com, cgudiol_ext@iconcologia.net, jcarratala@bellvitgehospital.cat, Infectious Diseases Department, Bellvitge University Hospital, University of Barcelona, Spain.

Objective: The usefulness of routine microbiological testing to propitiate antimicrobial de-escalation in hospitalized patients with community-acquired pneumonia (CAP) continues to be a subject of debate. We aim to determine the impact of clinical stability on the effect of de-escalation on 30-day mortality in patients with community-acquired pneumonia.

Methods: Data came from a prospective cohort of adults hospitalized for community-acquired pneumonia. A propensity score matching analysis was conducted for bias reduction in the comparison of de-escalation to a non-randomized control group. The role of clinical stability as a confounder of the effect of de-escalation on 30-day mortality was performed in the matched sample using a logistic regression model.

Results: From the included 3677 patients, a 3 to 1 matching was performed, resulting in a total of 2107 patients, 1360 unmatched and 747 matched. 1538 (73%) were clinically stable and 1498 (71.1%) had positive microbiology. The crude 30-day mortality was 37/1360 (2.72%) for non-de-escalated patients and 8/747 (1.07%) for de-escalated patients. The estimated OR for de-escalation in the matched sample was 0.39 (95% CI: 0.17-0.79). After adjusting for clinical stability, the OR for de-escalation moved to a non-significant result 0.53 (95% CI: 0.23-1.1).

Conclusions: The effect of de-escalation on 30-day mortality on community-acquired pneumonia without proper adjustment for clinical stability to avoid confounding by indication results in negative bias.

Keywords: De-escalation, propensity score, bias.

Parallel session: Omic data analyses

Chairperson:

Stefano Cabras, Universidad Carlos III de Madrid (Spain)

Variable selection in the omics data setting: Sure Independence Screening coupled with elastic-net and its variants

Arce Domingo-Relloso,[1,2,3] Yang Feng,[4] Zulema Rodriguez-Hernandez,[1] Karin Haack,[5] Shelley A. Cole,[5] Ana Navas-Acien,[2] Maria Tellez-Plaza,[1] Jose D. Bermudez[3]

[1] Department of Chronic Diseases Epidemiology, National Center for Epidemiology, Madrid, Spain.

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

[3] Columbia University Mailman School of Public Health, New York, NY, USA.

[4] Department of Biostatistics, New York University, New York, NY, USA.

[5] Population Health Program, Texas Biomedical Research Institute, San Antonio, TX, USA

The statistical analysis of omics data as markers of health endpoints poses a great computational challenge given the ultra-high dimensional nature and frequent between-feature correlations. In this context, adaptive elastic-net (Aenet) regression, a dimensionality reduction regularization method, has theoretically shown to provide less biased parameter estimations, as well as better predictive accuracy, compared to other regularization methods such as LASSO and elastic-net. Alternatively, the SIS R package combines the Iterative Sure Independence Screening (ISIS) algorithm to regularization methods such as LASSO, Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP), to efficiently improve variable selection.

In this work, we extended the SIS R package by pairing the ISIS algorithm with elastic-net and Aenet. We subsequently used real epigenomic data from the Strong Heart Study to compare the performance of the ISIS-paired penalization methods to Bayesian shrinkage and frequentist regression methods on the evaluation of genome-wide human blood DNA methylation as a marker of body mass index and lung cancer incidence.

The effect estimates for selected features in common across different methods were overall similar. For continuous BMI, Aenet outperformed the other methods in prediction, but showed the highest computational time. For incident lung cancer, SCAD and MCP outperformed in prediction and computational efficiency, but led to undesirably minimal sizes of selected features sets. The SIS package extended with the novel ISIS-paired dimensionality reduction functions, which also incorporates a bootstrap-based method to calculate confidence intervals for the estimated regression coefficients, is publicly available in CRAN (https://github.com/statcodes/SIS).

Addressing missing genes in gene expression meta-analysis with DExMA

Juan Antonio Villatoro-García[1], *Jordi Martorell-Marugán*[2], *Daniel Toro-Domínguez*[3], *Yolanda Román-Montoya*[4], *Pedro Femia*[5], *Pedro Carmona-Sáez*[6],

[1]javillatoro@ugr.es, Dpt. of Statistics and Operational Research, University of Granada and Bioinformatics Unit, Centre for Genomics and Oncological Research (GENyO)

[2]jordi.martorell@genyo.es, Dpt. of Statistics and Operational Research, University of Granada and Bioinformatics Unit, Centre for Genomics and Oncological Research (GENyO)

[3]daniel.toro@genyo.es, Medical genomics, Centre for Genomics and Oncological (GENyO)

[4]yroman@ugr.es, Dpt. of Statistics and Operational Research, University of Granada

[5]pfemia@ugr.es, Dpt. of Statistics and Operational Research, University of Granada

[6]pcarmona@ugr.es, Dpt. of Statistics and Operational Research, University of Granada and Bioinformatics Unit, Centre for Genomics and Oncological Research (GENyO)

The information regarding gene expression studies deposited in public repositories has experienced an exponential growth in recent years. In this context, meta-analysis methods have become one of the most popular statistical techniques to obtain new scientific knowledge from this type of data, because they allow researchers to combine the results from different gene expression studies into a single common result.

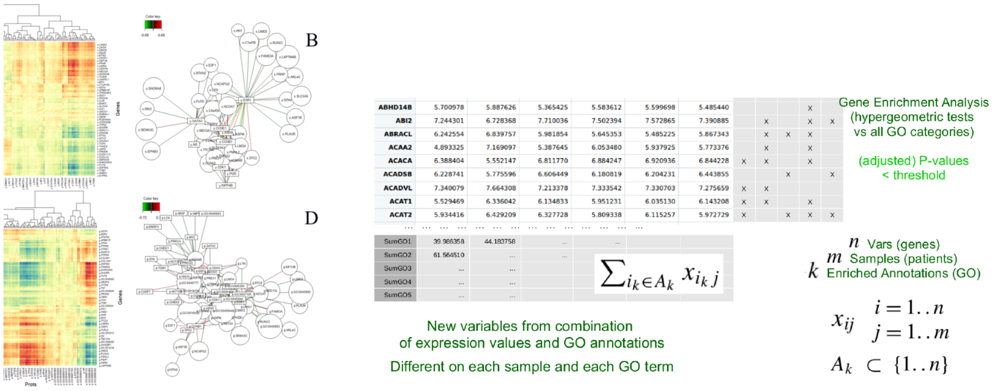
However, on several occasions, these methods have not been applied correctly, giving rise to inconsistent results and misinterpretations. One of the reasons why these errors occur is that there is a limited number of software that allows performing all the steps of gene expression meta-analysis. DExMA is an R package which implements all the necessary functions to properly accomplish gene expression meta-analysis. Moreover, DExMA contains functions that allow to control the possible existence of missing genes between the different datasets, as well as being able to carry out an imputation of them. The appearance of missing genes is a major problem that can arise when studies from different platforms are combined. DExMA is freely available in the Bioconductor repository: https://bioconductor.org/packages/DExMA/.

Keywords: missing values, imputation, meta-analysis

Integrative Analysis of Multi-Omics Data with Addition of Biological Knowledge

Ferran Briansó[1], *Alex Sánchez-Pla*[2]
[1]ferran.brianso@gmail.com, [2]asanchez@ub.edu
Department of Genetics, Microbiology and Statistics, University of Barcelona

Integrative analysis of multiple omics data allows, not only for the combination of heterogeneous data, but also for the combined use of biological data to extract information that could not be unveiled by the separated analysis of each of the original data types [Gomez-Cabrero, 2014]. One common approach to omics integration is using dimension reduction methods, which are also helpful for data visualization [Meng, 2016]. There is however one point that may be improved: the difficulty in interpreting results from the biological point of view [Yamada, 2021]. In the work presented here, biological annotations, such as GO Terms, Gene Sets or custom annotations, are combined with numerical values, such as protein or gene expression, using multiple factor analysis or related techniques, allowing to improve interpretability and providing better biomedical insights.

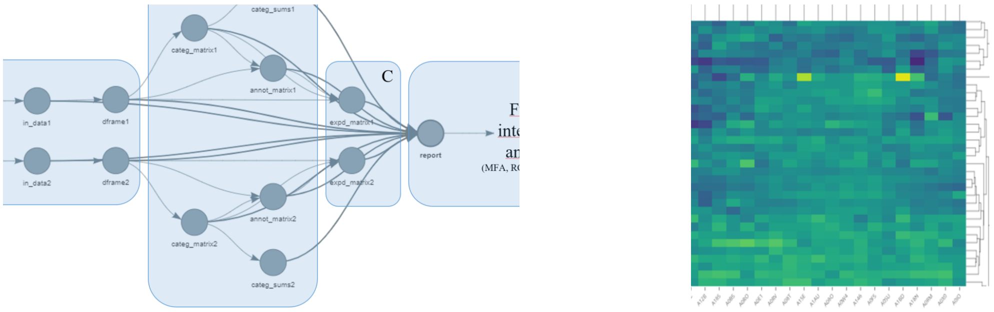


ABHD14B	5.700978	5.887626	5.365425	5.583612	5.599698	5.485440				x	
ABI2	7.244301	6.728368	7.710036	7.502394	7.572865	7.390885		x		x	x
ABRACL	6.242554	6.839757	5.981854	5.645353	5.485225	5.867343		x	x	x	
ACAA2	4.893325	7.169097	5.387645	6.053480	5.937925	5.773376		x		x	
ACACA	6.388404	5.552147	6.811770	6.884247	6.920936	6.844228	x	x		x	
ACADSB	6.228741	5.775596	6.606449	6.180819	6.204231	6.443855			x		x
ACADVL	7.340079	7.664308	7.213378	7.333542	7.330703	7.275659	x	x			
ACAT1	5.529469	6.336042	6.134833	5.951231	6.035130	6.143208	x	x		x	
ACAT2	5.934416	6.429209	6.327728	5.809338	6.115257	5.972729	x		x	x	x

SumGO1	39.986358	44.183758	...	...		
SumGO2	61.564510	...	...			
SumGO3	...	...				
SumGO4	...	...				
SumGO5	...	...				

Left: Some of the results of an analysis of 150 samples from TCGA. Heat maps (**A, C**) and association networks (**B, D**) resulting from the integration by Regularized Canonical Correlations Analysis with mixomics R package. Performed with the original data sets (**A, B**) or using data expanded with biological annotations to Gene Ontology (**C, D**), so adding some GO terms to the features from each source, where outputs contain higher level of information (higher density in both type of plots). **Right:** Representation of the process of integrating those annotated GO terms as new variables, new rows, as proposed by Busold et al. in 2005.

An R package with the methods developed, and a pipeline using the *targets* package, have been implemented to facilitate reproducibility and application to biomedical research.



Workflow of the steps implemented in the annotation and expansion of omics data: A couple of ‘omics-derived input data sets (e.g. pre-processed gene expression and protein abundance matrices) are converted to R data frames (**A**). Annotations are created, or loaded, as additional objects (**B**), one for each given input matrix, and used to expand these original data, to end up with a pair of data frames (**C**) containing the starting values plus the average expression/abundance values of the features related to each annotation as new variables. Finally, an R markdown report is rendered to show steps and main results of the process, and the output is used for further integrative analyses (**D**). Snapshot (**E**) of one of the heat maps created to show the expanded matrices resulting from the analysis.

Keywords: Multi-Omics; Integrative Analysis; Biological Annotations.

Exploring transfer learning in omics data

Ferran Reverter[1], *Esteban Vegas*[1], *Maxi Bless*[1]

[1]freverter@ub.edu, evegas@ub.edu, maxibless@gmail.com. Department of Genetics, Microbiology and Statistics, University of Barcelona.

The emergence of artificial intelligence has led to the implementation of deep learning models with a huge number of parameters that need to be estimated, which has promoted the need for the exchange and reuse of pre-trained models, the so-called *transfer learning*, which avoid the need to train the models from scratch and enables rapid development of new models from existing ones. In the scenario in which data are scarce, training a model from scratch might not be feasible. Instead, the model can be initialized with the majority of parameters from another model trained on a similar task. Transfer learning can be viewed as incorporating prior knowledge into the model. Parameters of the original model trained on a large data set are used for initialization for the target model trained on a related task but with much less data available. Since tasks are related, the transferred model may code features useful for the target task.

In the fields of computer vision and natural language processing, trained models are shared through repositories called model zoos and are available for popular machine learning frameworks (Keras, PyTorch, Tensorflow) but in omics there are not many platforms to share trained models.

We explore dense architectures for transfer learning in a common scenario in omics. In this field, often multiple and heterogeneous types of high dimensional data are available each of one having moderately large size but being small the subset of observations with complete data in two or more omics. Let us suppose we have data from two omic data types, types 1 and 2. Denoting by n_1 the number of observations in omic data type 1 and by n_2 , with $n_2 << n_1$, the subset of complete observations having both omics and, by p_1 and p_2 the number of variables in types 1 and 2, respectively. We investigated which configuration of values for the parameters (n_1, n_2, p_1, p_2) guarantee that the transfer of the model trained with data type 1 is useful to train predictive model for small size complete data. We used the overfit risk to measure the performance of the transfer learning.

Keywords: Transfer learning, Dense Neural Networks. Omics data.

AMS: 68T07, 68T09.

Parallel session: Covid-19

Chairperson:

Paloma Botella, Generalitat Valenciana (Spain)

COVID-19's incubation time period by vaccination status

Guadalupe Gómez Melis[1,2]*, Jordi Cortés*[1]*, Gabriela Abelenda-Alonso*[3]*, Alexander Rombauts*[3]*, Daewoo Pak*[4]*, Yu Shen*[5]*, Klaus Langohr*[1] *on behalf ot the DIVINE study project investigators.*

1. Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Barcelona, Spain
2. Institute of Mathematics of UPC - BarcelonaTech (IMTech), Barcelona, Spain
3. Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain
4. Division of Data Science, Yonsei University, Wonju, Korea
5. Department of Biostatistics, University of Texas, MD Anderson Cancer Center, Texas, US

Better understanding of the incubation period of an infectious disease is of public health and clinical importance. The finding is essential to estimate the reproductive number and assess how different public health interventions affect the outcome of the epidemic.

We have conducted an observational study based on a telephone survey by trained clinical personnel with the aim of estimating the COVID-19 incubation time, that is, the elapsed time from exposure to SARS-Cov2-infection to symptoms onset. Most prior studies used travel records from travellers from Wuhan during the first wave of the pandemic. However, data on later SARS-CoV-2 variations and vaccination status are scarce. The recruitment period corresponds to the delta driven 5th wave in Catalunya and started in July, 2021. Adult patients with hospital admission for COVID-19 were eligible. The main variables collected were the time or an interval of time relative to the infection and the symptoms onset. Other factors such as age, gender and vaccination status/type were also collected. We estimate the distribution of the incubation time period using the generalized odds-rate class of regression models accounting for the factors previously mentioned. The proposed method allows us to integrate all the patient information regarding infection time and symptoms onset time, even when incubation times are interval-censored or missing.

The final dataset collected between 6 July, 2021 and 13 December, 2021 includes 427 patients. Eleven patients without essential information were excluded. Among the remaining 416 patients, 46% had been fully vaccinated. In addition, 262 provided some information on the infection time and symptoms onset: 15% reported an exact incubation time, 48% provided an interval of dates for the infection time, and the remaining 37% only informed on the start of symptoms. The fifth wave of COVID-19 was associated with sustained community transmission and a high percentage of patients were unable to identify a specific risk exposure.

Our preliminary analysis suggests a shorter incubation period compared with the early variant during the delta driven fifth wave, and no differences stratifying by vaccination status, age, or sex. Incubation time estimates will be provided when the final dataset is validated. We will perform sensitivity analyses to contemplate several sources of bias such as recall bias as well as to assess the influence of the elapsed time after complete vaccination on the incubation time. As far as we know, our study is the first that estimates the incubation time based on the data including patients with mixed status of vaccination.

Keywords: COVID-19, Incubation period, Odds-rate models

1

COVIDCAT – Bayesian Small Area Estimation for the incidence of COVID-19 across Catalonia

Pau Satorra[1], *Marc Marí-Dell'Olmo*[2], *Aurelio Tobias*[3], *Cristian Tebé*[1]

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

[2] mmari@aspb.cat, Agència de Salut Pública de Barcelona (ASPB), CIBER Epidemiología y Salud Pública (CIBERESP), Institut d'Investigació Biomèdica (IIB Sant Pau)

[3]aurelio.tobias@idea.csic.es, Instituto de Diagnóstico Ambiental y Estudios del Agua (IDAEA), Consejo Superior de Investigaciones Científicas (CSIC)

Objective: To use bayesian small area estimation (SAE) to assess the incidence of covid-19 across the basic health areas (ABS) in Catalonia

Methods: Data came from the daily update of the Catalan open data COVID-19 registry by small primary care areas named basic health areas (ABS). Small Area Estimation (SAE) methods have been used to obtain reliable estimates of several indicators of the covid-19 incidence in ABS. We use Bayesian hierarchical models, applying the computational method Integrated Nested Laplace Approximation (INLA), to smooth the estimates of an area taking into account the results of the neighbouring areas.

Results: For each ABS the following indicators are estimated: smooth cumulative incidence rates, excess risk, probability of this excess risk being greater than zero and the probability of the incidence for the last 14 days being greater than some threshold. The analysis is performed by pandemic waves. Results are stratified by sex, and the trend of the last days is plotted for each indicator. Moreover, we build an R shiny application (https://ubidi.shinyapps.io/covidcat/) that automatically updates results every day. These results are presented in a fashionable way through maps and graphical representations to help make as much accessible as possible the comparison between ABS. Also, a Twitter bot posts every night several tweets with the basic results (https://twitter.com/covidcat4).

Conclusions: Bayesian SAE methods let us assess several indicators of the incidence of covid-19 into small areas taking in account the neighbouring areas, giving more reliable estimates. This application may help to trace the present and past state of all the different areas across Catalonia in an interactive and user-friendly manner, accessible to all public.

Keywords: Spatial Statistics, Small Area Estimation, Bayesian Statistics, Covid-19 surveillance.

Cluster analysis applied to COVID data: a study of the patient's characteristics and disease outcomes

Daniel Fernández[1], *Nuria Pérez-Alvarez*[2], *Gemma Molist*[3], *Erik Cobo*[4]

[1]daniel.fernandez.martinez@upc.edu, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Barcelona, Catalunya, Spain. Institute of Mathematics of UPC - BarcelonaTech (IMTech), Barcelona, Spain.

[2]nuria.perez@upc.edu, Fight AIDS and Infectious Diseases Foundation, HUGTIP, Barcelona, Spain. Department of Statistics and Operations Research, Universitat Politecnica de Catalunya, Barcelona, Catalunya, Spain. Universitat Oberta de Catalunya, Barcelona, Catalunya, Spain.

[3]gmolist@idibell.cat, Statistical Unit, Institut d'Investigació Biomèdica de Bellvitge, Barcelona, Catalunya, Spain.

[4]erik.cobo@upc.edu , Department of Statistics and Operations Research, Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Barcelona, Catalunya, Spain.

The COVID-19 outbreak has brought great challenges to healthcare resources around the world. Patients with COVID-19 exhibit a broad spectrum of clinical characteristics and disease outcomes. The purpose of this work was to determine profiles of COVID-19 patients with heterogeneous features over each of the first 3 waves (from October 2020 to February 2021).

The dimension of the data set was reduced with the aim of keeping the maximum information via performing classification trees and selecting the variables that classify better into different COVID outcomes. Furthermore, the variables with more than 20% of missing observations were removed.

A cluster analysis was performed on an anonymized database of 3381 patients and their information contained in a selected subset of the available 290 variables considered "important". Dissimilarity- and partitioning-based clustering methods (k-means, k-prototypes, and CLARA, among others) and likelihood-based methods (mixture models and Kamila, among others) for continuous, categorical, and mixed-type data. Internal validation of the clusters was carried out, finding a low level of grouping which remains clinically informative.

The identified clusters contain useful information to guide other studies to characterize groups of patients with this disease that could have a differential therapeutic response or prognosis.

Keywords: COVID, cluster, machine learning.

AMS: 62 Statistics; 92 Biology and other natural sciences.

Spatio-temporal small area surveillance of the COVID19 in Comunitat Valenciana

Botella-Rocamora, P[1], *Perez-Panadés, J*[2], *Martínez-Beneito MA*[3]

[1]botella_pal@gva.es, [2]perez_jorpan@gva.es, Dir.Gral.de Salut Pública i Adiccions,
Conselleria de Sanitat i Salut Pública-Generalitat Valenciana

[2]Miguel.A.Martinez@uv.es, Departamento de Estadística e IO, Universitat de València

Comunitat Valenciana is geographically divided into 542 municipalities and 241 administrative health zones which belong to 24 health departments. For the health authorities of this region is very important to know the magnitude of the pandemic and its trend at any given time for each of these areas in order to make public health interventions more effective.

The Covid19 pandemic has been a challenge in many ways, also for the traditional epidemiological surveillance tools. In this context, the development of methodologies that allow monitoring the evolution of the pandemic in areas as small as possible has been a priority. For this goal, spatio-temporal disease mapping models have proved to be a valuable tool of evident interest. Regretfully, these models were not originally devised for this pandemic context and the calculation of epidemic quantities like the instantaneous reproduction number Rt, of widespread use throughout this whole pandemic period, is not easily reproducible within these models.

In this work we introduce a spatio-temporal spline model particulary tailored for COVID-19 surveillance, which has allowed estimating and monitoring the Rt for small areas in Comunitat Valenciana. The results of this model have been weekly used by the health authorities of this region to nowcast the magnitude and trend of the COVID-19 rates at different small area levels. This tool has made it possible to identify the regions with the worst epidemic situation for each week, facilitating the surveillance work of the pandemic.

Keywords: disease mapping, surveillance, covid19.

AMS: AMS Classification (Optional).

1

Dynamic evaluation of COVID-19 clinical states and their prognostic factors to improve the intra-hospital patient management

C. Tebé Cordomí[1] *and G. Gómez Melis*[2] *on behalf ot the DIVINE study project investigators*

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

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

Objective: To achieve a deeper knowledge of the severe form of the disease caused by the SARS-CoV-2 virus a research project has been designed with multiples goals. First, to estimate the COVID-19 incubation period. Second, to identify the most clinically relevant prognostic factors for severe pneumonia, invasive respiratory support, or death in COVID-19 hospitalized subjects. Third, to develop and validate a prediction tool for severe pneumonia, invasive respiratory support, or death in COVID-19 hospitalized subjects. And fourth, to compare and cluster clinical COVID-19 hospitalized subjects' profiles between fourth waves of the pandemic.

Methods: Data came from a multicenter cohort study of hospitalized adults with confirmed COVID-19 in eight five hospitals in the Barcelona metropolitan south region. About 5,000 subjects from the 1st, 2nd, 3d, and 5th waves of the pandemic were included. A multidisciplinary research team, under the acronym DIVINE, integrated by researchers from the GRBIO (UPC-UB), Bellvitge University Hospital, and Bellvitge Biomedical Research Institute collaborated to define a statistical framework with a clear clinician focus to achieve the above-described goals.

Results: A generalized odds-rate class of regression models was used to estimate the distribution of the incubation time accounting for age, gender, and vaccination type, and considering the interval-censored nature of the incubation times. A semi-Markov multi-state model was used to identify the most relevant prognostic factors for a given wave and to estimate the probability of attaining different stages of disease progression within a given period, conditionally on a set of subject features. Finally, a clustering analysis was conducted to identify relevant hidden and relevant patient profiles. The results of this work will help clinicians to early identify patients at higher risk of suffering severe pneumonia or death.

Conclusions: The statistical framework used in this project will be susceptible of being applied in future pandemics with a similar disease course.

Keywords: Multi-state, generalized odds-rate class of regression model, Covid-19.

Parallel session: Biostatistical methods

Chairperson:

Dae-Jin Lee, Basque Center for Applied Mathematics (Spain)

Current statistical issues in platform trials for the evaluation of multiple treatments

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

[1]Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna
*marta.bofillroig@meduniwien.ac.at

Adaptive platform trials provide a framework to simultaneously study multiple treatments in a disease. They are multi-armed trials where interventions can enter as they become available and leave earlier for futility or efficacy based on interim analysis or after evaluation in the final analysis. The attractiveness of platform trials compared to separate parallel-group trials is not only due to operational aspects as a joint trial infrastructure and more efficient patient recruitment, also results from the possibility to share control groups, to efficiently prune non-efficacious treatments, and to allow for direct comparisons between experimental treatment arms. However, the flexibility of the framework also comes with challenges for statistical inference and interpretation of trial results. The major challenges concern the adaptivity of platform trials (decisions on adding or dropping arms cannot be fully pre-specified but may have an impact on current trial arms), multiplicity issues (due to multiple interventions, subgroups and interim analysis) and the use of shared controls (non-concurrent controls and blinding for interventions that have different routes of administration, change in the standard of care). We will overview some statistical approaches recently proposed to address these problems. In addition, we will give examples of disease-specific platform designs under investigation in the IMI project EU-PEARL (Grant Agreement no. 853966).

Keywords: Clinical trials, Statistical Inference, Multiple testing

Modelling multiple seasonalities of NO_2 hourly pollution levels

M. Avila[1], *A.M. Alonso*[2], *D. Peña*[3]

[1]matiasluis.avila@alumnos.uc3m.es, Department of Statistics, University Carlos III of Madrid
[2]amalonso@est-econ.uc3m.es, Department of Statistics and Institute Flores de Lemus
[3]daniel.pena@uc3m.es, Department of Statistics and UC3M-Santander Big Data Institute

NO_2 is one of the most common pollutants in urban areas and road traffic is the main source of this contaminant. The NO_2 hourly time series has three seasonal patterns due to human activity and climatological conditions. The classical approach assumes that those seasonalities are determinist and can be modelled by trigonometric functions or dummy variables. This assumption may be too strict. A more flexible model is to allow the seasonality to slowly change as in a seasonal ARIMA model, where the seasonality is modelled as a stochastic processes. In this paper, we propose to model them iteratively combining different seasonal ARIMA models.

We proposed a model that takes into account the regular dependency between hourly observations, the three seasonal components (daily, weekly and annual with seasonality 24, 168 and 8736 hours, respectively) and covariables such as the current average wind speed. It is worth noting that this model is not linear since the seasonal component is composed by varying parameters depending on the day hour and weekday. In order to estimate the model we have performed an approximation in two sequential steps: (1) In the first step we stratify the hourly NO_2 time series into 168 weekly time series, formed by each of the hours of the day and the days of the week. Each of the 168 weekly subseries is modelled separately with a seasonal $ARIMAX_{52}$ and covariables. The regular component of this $ARIMAX_{52}$ model captures the weekly seasonality while the seasonal one captures the annual seasonality. (2) Secondly, we consider the residuals from the first step in their natural order and fit a seasonal $ARIMA_{24}$. The seasonal component of this $ARIMA_{24}$ model will capture de daily seasonality while the regular component will capture the dependency between an observation and the immediately preceding ones. We compare our approach with other methods that have been developed to consider more seasonalities such as TBATS and Prophet, where the seasonal components are modelled by trigonometric functions.

Keywords: Pollution, Time Series, Multiple Seasonalities

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

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

[1]Dept. de Matemàtiques, UAB
[2]Dept. de Psicología Experimental, UCM
[3]Dept. de Psicología, UJA

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. The results are compared with a group of young subjects (Biology students). 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 50 head-tail outcomes, simulating the behaviour of a fair coin, without seeing (but perhaps remembering) the past outcomes. The study presented here is framed in the context of longitudinal data analysis in which a binary response, for a same individual, is repeated at 50 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). However, in activities concerning the generation of random values by humans, it seems unrealistic to assume that all the series are mentally produced according to a same order-memory. We propose latent class models based on Markov chains to classify individuals into different classes which differ in the form of generating head-tail sequences. We assume that subjects in the sample generate head-tail series according to a Markov chain, however, the order-memory of the chain varies across unobserved subgroups (classes). We analyse whether there are significant differences in the binary sequences mentally generated by healthy older adults compared to group of students.

Keywords: latent class model, higher-order Markov chains, random series production

References

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

Classification in Semiparametric Nonlinear Mixed Models using P-Splines and the SAEM Algorithm

Maritza Márquez[1], *Cristian Meza* [2], *Dae-Jin Lee* [3], *Rolando De la Cruz* [4]

[1]maritza.marquez@postgrado.uv.cl, CIMFAV - Faculty of Engineering, University of Valparíso
[2]cristian.meza@uv.cl, CIMFAV - Faculty of Engineering, University of Valparíso
[3]dlee@bcamath.org, BCAM - Basque Centre for Applied Mathematics
[4]rolando.delacruz@uai.cl, Faculty of Engineering and Sciences, University Adolfo Ibañez

Several authors, such as Marshall and Barón (2000), Arribas-Gil, et al. (2015) and De la Cruz, et al. (2017), analyzed a dataset from a clinical study related to the risk of loss for a group of pregnant Chilean women. At the time of delivery, each woman was classified into two groups: a normal group, in which those women who had a normal delivery were considered; and an abnormal group, in which only those women who had some type of delivery were taken into account, a complication that would result in a non-terminal delivery together with the loss of the fetus. In particular, in these previous works, the authors modeled the concentration of the hormone β-HCG in 173 women during the first trimester of pregnancy using mixed models. For the purposes of our research, we decided to extend these previous works, proposing **(i)** a nonlinear logistic mixed model with three random effects (NLME$_{(3)}$), which distinguishes it from previous works, since they only used one random effect (NLME$_{(1)}$) and **(ii)** an additive nonlinear mixed-effects model with spline penalty (SPNLME). Both models were estimated using the SAEM algorithm (Delyon et al. 1999), which is a stochastic approximation of the EM algorithm. On the other hand, we carry out a classification technique using the non-linear mixed models proposed via Importance Sampling, seeking to minimize the probability of misclassifying and point out the group to which each individual belongs.

Based on this proposal and on the progress made so far, we have managed to obtain improvements with respect to the adjustment of the data for the SPNLME additive model, also called semi-parametric, by including new spline bases with penalties. In addition, we implement the classification method on this semi-parametric mixed models, obtaining improvements with respect to the confusion matrix.

Group	**NLME$_{(1)}$**		**NLME$_{(3)}$**		**SPNLME**		**Total**
	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Total (173)
Normal	118	6	120	4	**122**	**2**	124
Abnormal	21	28	20	29	**19**	**30**	49
AUC	0.859		0.866		**0.887**		

Table 1: Classification of the $NLME_{(3)}$ and SPNLME models via Importance Sampling using the SAEM algorithm. The results are compared with the model $NLME_{(1)}$.

Keywords: Longitudinal data, classification, nonlinear mixed models.

Density regression via dependent Dirichlet process mixtures and penalised splines

María Xosé Rodríguez-Álvarez[1], *Vanda Inácio*[2]

[1]mxrodriguez@uvigo.es, Department of Statistics and Operations Research, Universidade de Vigo, Spain

[2]Vanda.Inacio@ed.ac.uk, School of Mathematics, University of Edinburgh, Scotland, UK

In many real-life applications, it is of interest to study how the distribution of a (continuous) response variable changes with covariates. Dependent Dirichlet process (DDP) mixture of normals models, a Bayesian nonparametric method, successfully addresses such goal. The approach of considering covariate independent mixture weights, also known as the single-weights dependent Dirichlet process mixture model, is very popular due to its computational convenience but can have limited flexibility in practice. To overcome the lack of flexibility, but retaining the computational tractability, this work develops a single-weights DDP mixture of normals model, where the components' means are modelled using Bayesian penalised splines (P-splines). We coin our approach as psDDP. A practically important feature of psDDP models is that all parameters have conjugate full conditional distributions thus leading to straightforward Gibbs sampling. In addition, they allow the effect associated with each covariate to be learned automatically from the data. The validity of our approach is supported by simulations and applied to two real datasets, one concerning the study of the association of a toxic metabolite on preterm birth, and the other one in the context of field trials experiments.

Keywords: dependent Dirichlet process, penalised splines, nonparametric regression

Parallel session: Software

Chairperson:

Natalia Vilor-Tejedor Beta Brain Research Center (Spain)

1

Let's MAMBO; Multivariate Analysis and Modelling of multiple Brain Outcomes in neurogenetic studies

Natalia Vilor-Tejedor, Blanca Rodríguez-Fernández, Patricia Genius, Diego Garrido-Martín, Roderic Guigo, Juan Domingo Gispert*

**nvilor@barcelonabeta.org,*

Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation, Barcelona, Spain; Centre for Genomic Regulation, The Barcelona Institute for Science and Technology, Barcelona, Spain; Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, Netherlands; Universitat Pompeu Fabra, Barcelona, Spain

Background: Neurogenetic studies aim to test how genetic information influences brain structure and function by combining neuroimaging-based brain features and genetic data from the same individual. Most studies focus on associations of individual genetic variants and single measurements of the brain. Despite the great success of univariate approaches, given the capacity of neuroimaging methods to provide a multiplicity of cerebral phenotypes, the development and application of multivariate methods become crucial.

Objective: In this work, several multivariate methods for modelling multiple brain outcomes and assessing its genetic influences were compared in terms of statistical power and performance over various levels and combinations of parameters.

Methods: We used a subset of publicly available data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database with available genetics and neuroimaging information. We also discussed relevant aspects of multi-trait modelling in the context of neurogenetic studies.

Conclusions: Multivariate methods have substantial potential to improve the predictive ability for brain outcomes in neurogenetic studies.

Keywords: Genetic association studies, Multivariate analysis,

AMS: 92D30

clustglm and clustord: R packages for clustering with covariates for binary, count, and ordinal data

Daniel Fernández[1], *Louise F. McMillan*[2], *Shirley Pledger*[2], *Richard Arnold*[2], *Ivy Liu*[2], *Murray Efford*[3]

[1]daniel.fernandez.martinez@upc.edu, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Barcelona, Spain.
[2]School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand.
[3]Department of Statistics, University of Otago, Dunedin, New Zealand.

We present two R packages for model-based clustering with covariates. Both packages can perform clustering and biclustering (clustering sites and species simultaneously, for example). Both use likelihood-based methods for clustering, which enables users to compare models using AIC and BIC as measures of the relative goodness of fit. The package *clustglm* implements techniques from Pledger and Arnold (2014) for handling binary and count data, or data from other single-parameter exponential family distributions, such as normal distributions with constant variance. It leverages *glm* and can fit pattern detection models that include individual-level effects alongside cluster effects. For example, when applied to presence/absence data, you can cluster sites and species while also taking into account any single-species effects, and any additional covariates. The package *clustglm* can also be applied to medical data, to cluster patients based on their response patterns over multiple questions. This package can accommodate balanced and nonbalanced designs, and numerical or categorical covariates. It provides the clustering equivalent of biplots, and also profile plots. We will illustrate the use of *clustglm* with a selection of ecological or medical datasets. The package *clustord* handles ordinal categorical data, using techniques outlined in Fernández et al. (2016). It builds on the ordered stereotype model, which accommodates flexibility in the ordinal scale used. The clustering results can reveal when two ordinal categories are effectively equivalent and can be combined to simplify the model.

Pledger, S. & Arnold, R. (2014). Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. Computational Statistics & Data Analysis, 71, 241-261.
Fernández, D., Arnold, R., & Pledger, S. (2016). Mixture-based clustering for the ordered stereotype model. Computational Statistics & Data Analysis, 93, 46-75.

Keywords: clustering, mixture models, multivariate, software.

AMS: 62H30, 62-04

A spatial epidemiological model for plant pathogen diseases spread

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

[1]cendoya_marmar@gva.es, Centre de Protecció Vegetal i Biotecnología, Institut Valencià d'Investigacions Agràries (IVIA)

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

An epidemiological model for almond leaf scorch caused by the phytopathogenic bacterium *Xylella fastidiosa* in Alicante was implemented at individual (tree) level. As data availability is scarce, the georeferenced distribution of trees was generated from the SIGPAC ("Sistema de Información Geográfica de Parcelas Agrícolas") database. In particular, following a planting frame of 7x7 m, our landscape consisted of 366668 trees in an area of about 1325 km^2.

In the proposed compartmental, individual-based model, the spatial relationship was introduced using the Matérn correlation function. Individuals were categorized according to their status: susceptible, asymptomatic infected and symptomatic infected. Whether a susceptible tree becomes infected depends on the strength of infection of infected individuals, given by the rate of infection parameter and correlation, whereas the appearance of symptoms is only time-dependent. Asymptomatic infected were considered infectious but with a reduced rate of infection compared to those with symptoms. We tested by simulation the behavior of the model parameters and the type of initial introduction (random or aggregated) on disease progression.

The greatest variability in results depended on the spatial range parameter and type of introduction. In areas with greater host continuity, i.e., without large empty holes, the rate of spread was nearly constant and depended on range. However, the regions without trees were a barrier to spread when their extent was greater than the range. Furthermore, to observe the effect of long-distance spread, which can be produced for example by the movement of plant material, new infected individuals were randomly introduced every few years. Simulations were implemented in `Python` due to its computational efficiency handling large databases. In order to interactively visualize the results of all scenarios, a `Shiny` application was also developed using `R`.

It is worth to note that this work could also be applied to the study of the spread of other pathogens by adapting the parameters involved in the model. It could also be very useful for establishing areas of higher risk as well as for the development of control strategies.

Keywords: Compartmental models, simulation, Matérn correlation

Managing REDCap Data: The R package REDCapDM

João Carmezim[1], *Judith Peñafiel*[1], *Pau Satorra*[1], *Esther García-Lerma*[1], *Natalia Pallarés*[1], *Cristian Tebé*[1]

[1] jcarmezim@idibell.cat, jpenafiel@idibell.cat, psatorra@idibell.cat, egarcia@idibell.cat, npallares@idibell.cat, ctebe@idibell.cat, Biostatistics Unit, Bellvitge Biomedical Research Institute (IDIBELL)

OBJECTIVES: The R package 'REDCapDM' is aimed to process REDCap data (Research Electronic Data CAPture) and to provide useful tools to carry out any task involved in the data cleansing process previous to data analysis.

METHODS: The R package 'REDCapDM' has several functions structured in three dimensions. First, read and process raw data from REDCap or through a REDCap API connection in R. Second, identify queries, concretely major errors, duplicated values, outliers, and missing values in data from REDCap in R. Third, make an automatic control of queries already resolved or pending resolution.

RESULTS: A total of 6 functions have been included in this first version of the package to accelerate and automate the process of data managing of REDCap data in R. The "redcap_data" function is used to read and prepare the data set originated from REDCap. The "rd_duplicates", "rd_expression", "rd_missings", and "rd_missings_dep" functions allow the identification of several types of data issues and generate a query report in several formats (xlsx, word, html, or pdf). In the query report, 6 columns are available: the record identifier, REDCap's instrument, type of query (e.g., missing value), description of the query, REDCap's event, and REDCap's repeat instrument. These columns are meant to help the user identify exactly the unwanted data point and correct/remove it from the data set. Finally, we have the "check_queries" function used to check the data set as many times as needed and identify which of the queries were already resolved.

CONCLUSIONS: The main advantage of using the package 'REDCapDM' is indeed the amount of time saved in the data cleaning process of data analysis that normally takes most of our time as well as having a single report with all types of queries in it. The package 'REDCapDM' is in the development phase and is not still available at the CRAN library.
Keywords: REDCap, data manage, quality data

Space-time interactions in Bayesian disease mapping with NIMBLE

A. Urdangarin[1], *T. Goicoa*[2], *M.D. Ugarte*[3]

[1]arantxa.urdangarin@unavarra.es, Department of Statistics, Computer Science and Mathematics,
Public University of Navarre

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

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

Spatio-temporal disease mapping models are extensively used to smooth risks and hence reduce variability. These models present some identifiability issues and specific constraints are needed to correctly identify and interpret the different terms in the model. In this work we will focus on two of the recent free software packages to fit spatio-temporal models in disease mapping within a fully Bayesian approach: NIMBLE and R-INLA. NIMBLE is a recent tool that permits to fit hierarchical models using Markov chain Monte Carlo (MCMC) algorithms. R-INLA is an R-package for approximate Bayesian inference based on integrated nested Laplace approximations and numerical integration.

The main goal of this work is to show how to fit spatio-temporal disease mapping models in NIMBLE. An Intrinsic conditional autorregresive (ICAR) prior distribution for the spatial random effects, a first order random walk (RW1) for the temporal random effects and four types of space-time interactions are assumed. To overcome the identifiability issues, we take advantage of one construction in NIMBLE to fit intrinsic conditional autoregressive models that constraints the random effects to sum to zero. This requires to express the interaction term as independent standard normal random variables and to include the spatial and temporal dependence through pre and post multiplication by appropriate matrices. Breast cancer mortality data in Spain during the period 1990-2011 is used for illustration purposes and a simulation study is conducted to compare results produced by both packages NIMBLE and R-INLA. Both procedures provide identical estimates of parameters and relative risks. The simulation study verifies that R-INLA and NIMBLE estimates of parameters and relative risks are close to the real values. Finally, in terms of computational time, R-INLA is quite faster than NIMBLE.

Keywords: NIMBLE, spatio-temporal interactions, sum-to-zero constraints

Poster session

Chairperson:

M. E. Castellanos, URJC (Spain) and V. Gomez Rubio, UCLM (Spain)

1

A comparative study of growth curves modelling approaches for the estimation of the age at peak height velocity

María Alejandra Hernàndez Velandia[1] *Dae-Jin Lee*[2], *María Xosé Rodríguez Álvarez*[3]

[1]mhernandez@bcamath.org, BCAM-Basque Center for Applied Mathematics,
Departament of mathematics, UPV/EHU - Universidad del País Vasco

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

[3]mxrodriguez@uvigo.es, Department of Statistics and Operations Research, Universidade de Vigo

The knowledge about growth curves has been of interest to different scientific fields, and it has led to the development of different approaches to the modelling of growth data. These models allow estimating growth trajectories as well as the study of maturational status and timing. For instance, it is often of interest to estimate the chronological age at which peak height velocity (aPHV) is reached and the rate of growth that occurs during peak height velocity (PHV) during puberty.

On the other hand, in clinical practice, left hand and wrist radiography-based bone age assessments such as Bayley–Pinneau (BP) or Tanner-Whitehouse (TW) methods are widely used to predict heights based on the fact that bone age correlates with a specific percentage of mature height reached in a specific moment when chronological age is constant.

In this work, we investigate different methods to estimate the growth in children and its derivative curves to obtain the aPHV. The advantages and disadvantages of the different growth curve models are illustrated with an example of heights of children aged between 10 and 18 years where the information of bone assessment before the age of 14 of the BP and TW methods are available.

Keywords: Growth curves, derivative curves, age at peak height velocity.

White blood cell profiles in Myotonic Dystrophy Type 1 mice overexpressing the *MSI2* gene.

Carlos J. Peña[1], *Maria Sabater-Arcis*[2], *Nerea Moreno*[2], *Juan A. Carbonell*[1]

[1]Unidad de Bioinformática y Bioestadística. Instituto de Investigación Sanitaria (INCLIVA)
[2]Instituto Universitario de Biotecnología y Biomedicina, Universitat de València
Corresponding author: jacarbonell@incliva.es

Background: Myotonic dystrophy type 1 is a rare disease that affects muscles and other body systems. Recent studies relate this disease to an overexpression of MSI2 gene. Our primary aim is to evaluate the effect of the expression of this gene on relative white blood cell counts.

Materials: Data for this study consist of 35 observations of proportions of each white blood cell in blood samples from mice that belong to four different treatment groups: healthy mice (FVB), diseased mice treated with a buffer solution (PBS), diseased mice treated with an empty virus (DES) and diseased mice treated with a virus overexpressing the gene of interest (MSI).

Methods: We begin by expressing our compositional data set in isometric log-ratio coordinates (*ilr*). Afterwards, a principal component analysis (PCA) is performed and the resulting loadings and scores are back-transformed to the centered log-ratio space where the compositional biplot can be shown in order to uncover relations in our compositional data set. This is then contrasted by plotting ternary diagrams.
Additionally, a linear discriminant analysis (LDA) is applied to our *ilr*-transformed data set in order to find the best way of separating the treatment groups using functions of the available variables.
Lastly, a Multivariate Analysis of Variance (MANOVA) is used for studying the effect of treatment on the whole white blood cell compositions and Tukey HSD tests are conducted in order to identify which treatment groups contributed to a significant global effect.

Results: Discriminant analysis suggests that the MSI group of diseased mice is clustered on the opposite side of the rest of the groups. Additionally, the MANOVA contrast indicates that statistically significant difference was found in the white blood cell compositions among the groups. Nevertheless, Tukey HSD *post-hoc* tests confirm that these differences are not detected between healthy and diseased mice but, instead, between MSI and PBS groups.

Conclusions: Classical statistical analysis techniques assume that the data can exist anywhere in real space. Therefore, they should not be performed without acknowledging the compositional nature of the data.

Keywords: compositional data, white blood cells, percentages

Estimating essential habitats combining fishery-dependent and -independent data applying Bayesian learning

Mario Figueira[1], *Xavier Barber*[2], *David Conesa*[1], *Antonio López-Quílez*[1], *Joaquin Martinez-Minaya*[3], *Pennino Maria Grazia*[4], *Iosu Paradinas*[5]

[1]Dpt. Estadística i Investigació Operativa, Universitat de València
[2]Centro de Investigación Operativa, Universidad Miguel Hernández de Elche
[3]Estadística e Investigación Operativa, Universidad Politécnica de València
[4]Centro Oceanográfico de Vigo, Instituto Español de Oceanografía, Vigo
[5]Scottish Oceans Institute, University of St Andrews, St Andrews

Species mapping is an essential tool for conservation managers as it provides a clear picture of the distribution of marine resources. However, in fishery ecology, the amount of objective scientific information is limited and data may not always be directly comparable or easily integrated. Information about the distribution of marine species can be derived from two main sources: fishery-independent data (scientific surveys at sea) and fishery-dependent data (collection and sampling by observers in commercial vessels). The aim of this study is to propose a novel approach to integrate these two different sources through sub-sequential Bayesian models. In particular, we apply the Bayesian learning paradigm in a sequential way by using as prior information for the second model that obtained as posterior distribution of the parameters of the first model, providing so better estimations and predictions. We compare results of both possible directions: updating dependent data with that obtained with independent data and vice versa) with a simulated example and a real case study using data from two elasmobranch species (S. canicula and G. melastomus) in the Mediterranean Sea.

Keywords: Spatio-temporal modelling, fisheries, Bayesian learning

AMS: 62P12, 62L10

1

New developments on Integral Priors for Bayesian Model Selection

D. Salmerón Martínez[1]*, J. A. Cano Sánchez*[2] *and C. P. Robert*[3]
[1]dsm@um.es, Departamento de Ciencias Sociosanitarias, Universidad de Murcia
[2]jacano@um.es, Departamento de Estadística e Investigación Operativa,
Universidad de Murcia
[3]xian@ceremade.dauphine.fr, Université Paris-Dauphine

Integral priors were developed for Bayesian model selection and have been successfully applied in many situations. However, there are two aspects that deserve special attention. First, the method is stated for the comparison of two models. Second, nonparametric density estimates of the integral priors have been typically needed to approximate the Bayes factors, which translates into more computing time. Here we generalize the definition for more than two models and propose new numerical procedures to approximate the Bayes factors. The method is illustrated with several examples including location-scale models, Poisson versus the negative binomial family, hypothesis testing for the exponential distribution mean, and the problem of testing if the mean of the normal distribution with unknown variance is negative, zero, or positive. Finally we illustrate the method for the variable selection problem.

Keywords: Integral priors, Bayesian model selection, Objective Bayes factor, Markov chains.

Exploring statistical methods for classifying individuals in extreme aging groups

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

Background

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

Methods

This study included 340 middle-aged cognitively unimpaired participants at risk of AD from the Alzheimer's and Families (ALFA) cohort. We used telomere length (TL) and magnetic resonance imaging (MRI) brain features for computing the aging metrics. Telomere length (TL) was determined by qPCR from DNA extracted from peripheral blood leukocytes. Delta-age from TL data was calculated as the residuals from regressing TL z-scores (normalized measures) on chronological age separately in women and men. Moreover, we used brain features to compute a BrainAge metric. The model of healthy brain aging was first trained with the chronological age and pre-processed structural MRI data of a training sample using a gradient boosting algorithm for capturing the multidimensional aging patterns throughout the whole brain. Then, BrainAge metrics were calculated on the testing sample as the difference between the estimated and chronological age using the already trained model. Positive metric values indicate accelerated aging, while negative metric values indicate decelerating aging. Accelerated and decelerated aging groups were defined as those 85 more extreme individuals of the computed metrics, roughly corresponding to the 10th and 90th percentile.

Results

Significant differences were found between aging groups and between sexes for both metrics, showing the classification ability of both estimated aging metrics (p-values < 0.001). None of the other demographic variables tested (chronological age, *APOE* ε4 carriership, years of education…) showed significant differences between the groups (p-values > 0.05).

Conclusions

The comparisons performed in this preliminary study confirmed the classification ability of the calculated age metrics between the groups. The classification ability was specific to these metrics, as we found no other significant effect when testing other demographic variables. Future analyses will focus on the analysis of circulating blood factors that differ between groups, integrating proteomics and metabolomics data to unravel the biological mechanisms associated with variability in the rate of aging.

Keywords: Aging, Bioinformatics, Classifying methods,

Dealing with non-proportional hazards: Impact of treatment strategies on mortality in Pseudomonas aeruginosa bacteraemic pneumonia in neutropenic cancer patients

Pallarès N[1]*, Albasanz-Puig A*[2]*, Gudiol C*[3]*, Tebé C*[1]

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

[2]aalbasanz@bellvitgehospital.cat. Infectious Diseases Department, Bellvitge University Hospital, IDIBELL, University of Barcelona, Spain.

[3]cgudiol_ext@iconcologia.net, Infectious Diseases Department, Bellvitge University Hospital, IDIBELL, University of Barcelona, Spain. Institut Català d'Oncologia (ICO), Hospital Duran i Reynals, IDIBELL, Barcelona, Spain.

Objective: To deal with non-proportional hazards in a Cox model to study treatment effect in mortality in patients with Pseudomonas aeruginosa (PA).

Methods: A multicenter, retrospective cohort study including neutropenic patients with PA bloodstream infection (BSI) was conducted in 34 centers (12 countries) between 2006 and 2018. Patients were classified into three groups according to the treatment received.

A Cox proportional hazards model was used to perform an adjusted analysis of the treatment strategies with age, sex, and septic shock as clinically relevant factors for mortality. The presence of septic shock at admission violated the proportional hazard assumption of the Cox model. To deal with this, a time-dependent covariates analysis on septic shock was performed. Data were split into three groups (from day admission to day 2, from day 2 to day 10, and from day 10 to day 30) defined after the exploration of the Schoenfeld residuals plot, and a clinical discussion on the rational of this cut points. In each time interval, a different coefficient was estimated for septic shock. Treatment effects were reported using hazard ratio (HR) and 95% confidence intervals.

Results: Among 1017 episodes of bacteremic pneumonia, PA was the source of BSI in 294 of them (31.3% women, 61.3-yo). 23.1% of patients received inappropriate initial empirical antibiotic treatment, 26.9% appropriate empirical combination and 50.0% appropriate empirical monotherapy. Almost half of them (48.1%) had septic shock at admission. Schoenfeld residuals showed non-proportional hazards for septic shock. After splitting data in three groups, appropriate empirical combination was a protective factor for mortality (HR 0.41, 95% CI 0.26-0.65). No differences were found between appropriate empirical monotherapy and inappropriate empirical treatment. Regarding septic shock, association with mortality was higher between admission and day 2 (HR 7.24, 95% CI 4.12-12.72), than between day 2 and day 10 (HR 3.47, 95% CI 1.98-6.09). After day 10, septic shock had no effect on mortality.

Conclusions: Splitting data into time intervals was a relevant strategy to solve the proportional hazards issue, and to give more information to clinicians about the weight of septic shock in early mortality and the effect of empirical antibiotic treatments.

Keywords: Infectious diseases, Non-proportional hazards, Cox model.

Optimal designs for a non-linear model of ethanol elimination in the human body

M.T. Santos Martín[1], *J.M. Rodríguez Díaz*[2], *I. Mariñas del Collado*[3]
[1] maysam@usal.es, Department of Statistics, University of Salamanca
[2] juanmrod@usal.es, Department of Statistics, University of Salamanca
[3] marinasirene@uniovi.es, Department of Statistics and Operational Research and Didactics of Mathematics, University of Oviedo.

Abstract:

The equation usually employed by forensic scientists to estimate the alcohol concentration in blood in a person after the ingestion of alcoholic drinks, has zero-order kinetics in the ethanol elimination phase, i.e., the elimination process occurs in the body at a uniform rate as a function of the ethyl-oxidation constant. The model, formulated by Widmark, does not consider the phase of increase in concentration, and approximates the phase of elimination in a linear way, which may be insufficient if the tests are carried out in the first phases of alcohol intake.

In this work, optimal designs are proposed for a non-linear model that fits the different phases of the pharmacokinetic process of ethanol in the human body (absorption, distribution, metabolism, and elimination) in order to compute the most informative observation times for the estimation of the parameters of the non-linear model. In the calculation of the designs, a covariance structure between responses in needed, since the parameters of the non-linear model depend, among other factors, on the characteristics of the subject in which the observations are taken.

Keywords: Optimal design of experiments, D-optimality, ethanol, Widmark

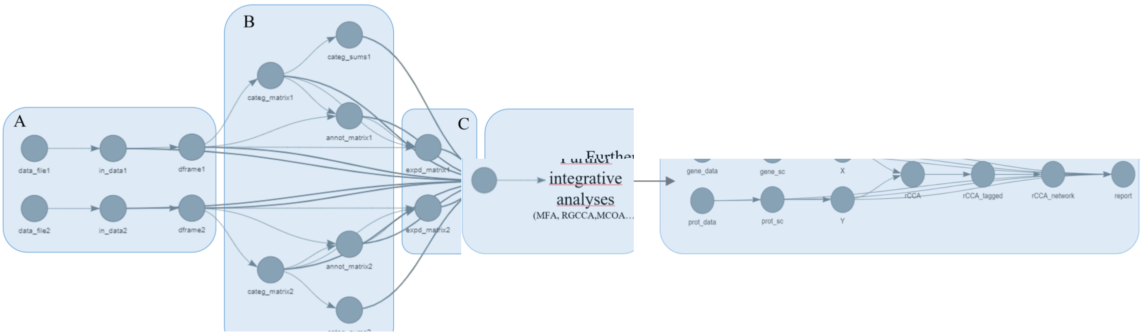
An R pipeline using the "*targets*" package for Multi-Omics Integrative Analyses

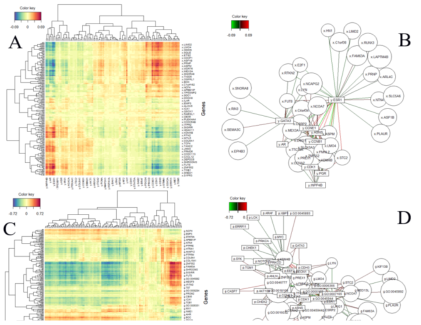
Ferran Briansó[1], *Alex Sánchez-Pla*[2]

[1]ferran.brianso@gmail.com, [2]asanchez@ub.edu

Department of Genetics, Microbiology and Statistics, University of Barcelona

Integrative analysis of multiple omics data allows, not only for the combination of heterogeneous data, but also for the combined use of biological data to extract information that could not be unveiled by the separated analysis of each of the original data types [Gomez-Cabrero, 2014]. One common approach to omics integration is using dimension reduction methods, which are also helpful for data visualization [Meng, 2016]. Several tools have been implemented as frameworks for the development of bioinformatic pipelines [Leipzig, 2017; Wratten, 2021], in most cases conforming complex workflow managers, such as Galaxy [usegalaxy.org] and NextFlow [www.nextflow.io]. These platforms, however, require a certain degree of technical knowledge in order to be configured, often away from the standard technical level of biomedical researchers and analysts. In the work presented here, we show a pipeline for the integrative analysis of multi-omics data, implemented with the *targets* package [books.ropensci.org/targets] that, unlike most pipeline toolkits, which are language agnostic or Python-focused, allows data scientists and researchers to work entirely within R. The example presented here performs an integrative multi-omics analysis, combining protein and gene expression data (from public sources such as TCGA) with biological annotations (GO, pathways or custom annotations), and applying distinct dimension reduction techniques (Regularized Canonical Correlations Analysis and Multiple Factor Analysis) implemented in proven reliable packages (mixOmics[mixomics.org/] and FactoMineR[factominer.free.fr/], respectively).





Above: *Targets* workflow of the steps implemented in the multi-omics integrative pipeline: A couple of 'omics-derived input data sets (e.g. pre-processed gene expression and protein abundance matrices) are converted to R data frames (**A**). Annotations are created, or loaded, as additional objects (**B**), one for each given input matrix, and used to expand these original data, to end up with a pair of data frames (**C**) containing the starting values plus the average expression/abundance values of the features related to each annotation as new variables. Then, a first R markdown report is rendered to show steps and main results of the annotate-and-expand process, and this output is used for further integrative analyses (**D**). In this case, the Regularized Canonical Correlations Analysis carried out with mixOmics (**E**) that ends up with a second markdown report.

Left: Some of the results of an analysis of 150 samples from TCGA. Heat maps (**A, C**) and association networks (**B, D**) resulting from the integration by RCCA with mixomics R package. Performed with the original data sets (**A, B**) or using data expanded with biological annotations to Gene Ontology (**C, D**), so adding some GO terms to the features from each source, where outputs contain higher level of information (higher density in both type of plots).

Keywords: Workflow Pipelines; Multi-Omics; Integrative Analysis.

Assessment of uncertainty in biomass estimation using different statistical tools employed in fisheries management

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

[1]alba.fuster1398@gmail.com, Master of Biostatistics student, Universitat de València
[2] Valencian Bayesian Research Group, Universitat de València
[3] Centro oceanográfico de Vigo - Instituto Español de Oceanografía (IEO - CSIC)

In recent decades, the concern about marine ecosystems health has led to a better management of fishery resources. However, these improvements have not been sufficient to reverse the global declining trend of overfished stocks. As a result, it turns necessary to improve the quality of scientific advice for fisheries management. One way to help managers is quantifying the unknown biomass of fish species. Estimating this biomass is not an easy task and so, several statistical models have been developed to describe that temporal behavior for those target species.

The aim of this study is to shed light and help scientists to choose amount statistical methods proposed in the literature for estimating the temporal behavior of the biomass. To do so, we simulate a biomass in a regular area and we set two different scenarios that reproduce the main data sources used to fit the statistical models:

- Oceanographic surveys (fishery-independent data), consisting of a georeferenced random sampling, where the information is given as relative biomass indices; and,
- Observers on fishing vessels (fishery-dependent data), consisting of a georeferenced preferential sampling, where the information is given as catch per unit effort (CPUE) indices.

We propose the use of R-INLA, a package in R that do approximate Bayesian inference for Latent Gaussian Models, in order to run different models (GLMs, GAMs and spatio-temporal models) for each of the different simulated scenarios. Results of the model provide us predictions of the CPUE and relative biomass indices. Finally, these predictions are fed into the SPiCT (stochastic surplus production model in continuous time) stock assessment model to estimate the "unknown" biomass. As the biomass has been simulated, we can assess the accuracy of this estimation using error measures as RMSE, MASE or MAPE and choose the best model.

Keywords: geostatistical models, simulation and accuracy

Comparison of three scoring methods applied to the ASES-p scale for shoulder pathology

Maider Mateo-Abad[1], *Kalliopi Vrotsou*[2]

[1]maider.mateoabad@biodonostia.org, Biodonostia, Primary Care Research Group; REDISECC
[2]kalliopi.vrotsoukanari@osakidetza.eus, Biodonostia, Primary Care Group; REDISECC

Shoulder pathologies are among the commonest musculoskeletal problems, and they are known to limit daily life activities, increase work absence and affect psychological and social well-being. One of the most implemented pathology-specific scale is the American Shoulder and Elbow Surgeons patient self-report section (ASES-p). The ASES-p is an 11-item scale which evaluates pain level and 10 activities of daily living. The scale is validated and it is reliable, but some studies indicated that the scoring of the items could be optimized. The optimization of the scoring or ponderation of the items could help to make better clinical decisions and to assess the effectiveness of an intervention, among others. Therefore, the aim of these study is to evaluate and compare three different scoring procedures, raw scoring, scoring based on factorial analysis (FA) and based on multiple correspondence analysis (MCA).

The Raw score was calculated by adding the patient's ratings. Factor loading were used to ponderate the items for the FA based score, and for the MCA based score, the coordinates of the main dimension were assigned to each category of all items. The scores were transformed to range from 0 to 100. Psychometric properties were evaluated for each scoring method, on a sample of 106 patients with shoulder pathologies. Reliability was tested with Cronbach's alpha and Spearman's correlations coefficients were obtained to analyze the convergent validity. Sensibility to change was explored comparing the scores of the patients before and after receiving a treatment.

Results were acceptable for the three scoring methods. Cronbach's alpha was higher for MCA score, than for FA and Raw scores (0.92 vs 0.90 and 0.87). The first also presented better internal consistency and better convergent capacity with the Constant-Murley score (0.67 vs 0.65 and 0.62), and with factors related with movement and strength. The raw score presented the highest correlation with pain related factor and the Short Form Health Survey. The FA based score shows the better sensibility to change, in all patients.The three methods showed good discriminatory capacity between known groups.

Keywords: Score, Multiple Correspondence Analysis, Factor analysis

Assessing geographical differences in the risk of recurrent hip fracture and death via Bayesian spatial illness-death models.

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

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

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

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

Illness-death models are a class of stochastic models, inside the multi-state framework, in which individuals are allowed to move over time between different states regarding illness and death. By means of transition probabilities they allow to study progression through different health conditions. We focus on spatial illness-death models with random effects for assessing correlation between spatial units. We apply this model to a cohort study of progression after osteoporotic hip fracture. Data come from the PREV2FO cohort, including patients aged 65 years and older who were discharged alive after hospitalization due to an osteoporotic hip fracture during 2008-2015 in the Valencia region, Spain. We assess geographical differences in the risk of recurrent hip fracture, death without refracture and death after refracture. We use a Bayesian approach using the integrated nested Laplace approximation (INLA).

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

Identification of severity prognostic factors in hospitalized patients with confirmed influenza by means of a multi-state model

Lesly Acosta[1], Mireia Jane[2,3,5], Nuria Torner[3], Luca Basile[2], Ana Martinez[2,3], Cristina Rius[3,4], Nuria Soldevila[3,5], Ángela Dominguez[3,5]

[1]lesly.acosta@upc.edu, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona-TECH
[2] Agència de Salut Pública de Catalunya, Barcelona, Spain
[3] CIBER Epidemiología y Salud Pública, Instituto de Salud Carlos III, Madrid, Spain
[4] Agència de Salut Pública de Barcelona, Barcelona, Spain
[5] Department de Medicina, Universitat de Barcelona, Spain

Multi-state models (MSM) are useful to study the evolution of a health problem that can be classified into different stages. It is characterized by its states and the transitions among them. Transition intensities are modelled as functions of covariates by means of separate proportional hazards models. This methodology assumes that future time course depends only on the present state, but not on previous history (Markov property).

In this work, we used a MSM to analyze disease course of severe hospitalized laboratory confirmed influenza (SHLCI) patients and to investigate factors associated with transitions between states during hospitalization. The main goal was the assessment of prognostic factors for severe outcomes in SHLCI.

The data was gathered from a retrospective cohort study of 1306 SHLCI cases registered by the 14 hospitals included in the Influenza Surveillance System of Catalonia (PIDIRAC) from 1 October 2017 to 22 May 2018. The MSM used to analyze the data consisted of three transient states (ward admission; ICU admission; return to ward), three absorbing states (discharged home; derived to a long-term care facility-LTCF; death), and nine transitions among these states. The covariates include comorbidities, age, gender, and vaccination status of the patients.

The results of the MSM will be very useful to provide insights for influenza preventive measures, the optimization of patient care and medical resources during peak influenza epidemic activity.

Keywords: Confirmed influenza hospitalization; Multi-state models; Markov property

1

A new algorithm to detect homogeneous damage zones applying unsupervised machine learning to remote sensing data

Leandro Sosa[1], *Ana Justel*[2], *Íñigo Molina*[1]

[1]leandroleonelsosa@gmail.com (L.S.); inigo.molina@upm.es (I.M.), Department of Topographic Engineering and Cartography, Universidad Politécnica de Madrid

[2]ana.justel@uam.es, Mathematics Department, Universidad Autónoma de Madrid

In the last decades, both the frequency and intensity of extreme weather and climate events have increased worldwide, leading to huge economic losses. Hailstorms usually result in total crop loss. Following a hailstorm, an insurance claims adjuster goes to the field and estimates the yield losses due to hail damage. The accuracy of estimates relies heavily on the detection of Homogeneous Damage Zones (HDZs), which allows extrapolating the data from field samples to the whole area of the damaged field. Currently, estimates of the surface area of HDZs are performed visually, and the adjuster has no information about either the affected area within the field or the degree of damage before getting to the site for inspection.

This research suggests developing an algorithm for automatic detection of HDZs through the application of unsupervised machine learning techniques to vegetation indices calculated from remote sensing data. Vegetation indices allow identifying and describing structural, phenological and biophysical parameters, which in a temporal line basis help detecting changes in vegetation. Five microwave and five spectral indices were evaluated before and after a hailstorm in zones with different degrees of damage. Dual Polarization SAR Vegetation Index and Normalized Pigment Chlorophyll Ratio Index were the most sensitive to hail-induced changes. The steps of the algorithm are: 1) selection and pre-processing of satellite data to remove speckle noise, presence of clouds and other atmospheric phenomena; 2) calculation of vegetation indices with highest sensitivity to hail-caused damage to construct the data matrix for pixel classification; 3) preparation of the variables for classification and data matrix construction; 4) unsupervised pixel classification into HDZs. The time series and rates of change of these indices were used as input variables in the K-means method for clustering pixels into homogeneous damage zones. To validate the algorithm, 38 storms that occurred between 2017 and 2020 were analyzed in 91 soybean, wheat and corn plots located in the Argentine Pampean plain. The One-Way ANOVA model ($p < 0.05$) was applied to each plot.

Validation of the algorithm showed that in 87.01% of cases there was significant evidence of differences in average damage between zones determined by the algorithm within the plot. Thus, the algorithm presented in this paper allowed efficient detection of homogeneous hail damage zones, which is expected to improve accuracy and transparency in the characterization of hailstorm events.

Keywords: Remote Sensing, Machine Learning, Agriculture.

Parallel session: Sport

Chairperson:

Irantzu Barrio, Universidad del País Vasco (Spain)

1

Closing the gender science and statistical research gap in sports medicine: The case of FC Barcelona, the current women European Champion league winner

Juan R Gonzalez[1,2]*, Eva Ferrer*[3]*, Guillermo Quintás*[4]*, Gil Rodas*[3,4]

[1]juanr.gonzalez@isglobal.org, Department of Bioinformatics, Barcelona Institute for Global Health (ISGlobal), [2]Department of Mathematics, UAB; [3] Football Club Barcelona Medical Department & Barça Innovation Hub; [4] Leitat Technological Center

There exists a large body of evidence describing why sex and gender should be considered in preclinical, clinical, and population research (Rich-Edwards, et al. 2018). However, there is an evident gap in favour of men over women in sport research performance. Therefore, there is an urgent need of research conducted on female athletes with the final aim of develop accurate evidence-informed approach to practice. This will overcome the problem of applying evidence developed in male athletes to female athletes that may be erroneous given biological, social and environmental differences among them (Emmonds, et al. 2019).

From a statistical point of view, data analyses should also consider existing gender differences. First, including different risk factors than those analysed in men. For instance, developing predictive models to predict risk injury in soccer players, should not only consider variables such as internal and external training load, clinical, physiological and nutritional factors, but also consider specific female risk factors such as body composition, hormonal or anatomical variables. Second, new statistical approaches (rather than stratified analyses) are also required to model female-specific factors such as menstrual patterns.

In this talk, we will introduce data from a cohort of 24 soccer elite players belonging to the professional women team of FC Barcelona who are being actively followed up from three consecutive seasons (2019-20, 2020-21, 2021-22). We have collected daily data from external training load data using electronic devices, non-contact injuries, genomic and longitudinal urine metabolomic data measured each 4 months. Additionally, we also have information about variables related to menstrual period and hormone profiles.

Properly predicting indirect (non-contact) muscle injuries are crucial to professional teams since they reduce lime lost to training or competition that, in turn, negatively impact financial resources and team performance. In this talk we will introduce new statistical models to predict risk injury (recurrent events) using high-dimensional data (devices), longitudinal metabolites, genomic (genomic variants and polygenic risk scores) and hormonal data.

Keywords: Gender gap, sportomics, risk injury prediction, metabolomics, longitudinal data.

References:

Rich-Edwards JW, et al (2018). Sex and Gender Differences Research Design for Basic, Clinical, and Population Studies: Essentials for Investigators. Endocrine Review, 39;4:424-439.

Emmonds S, Heyward and Jones B (2019). The Challenge of Applying and Undertaking Research in Female Sport. Sports Medicine - Open, 5, article number 51.

A hidden Markov model for assessing the hot hand phenomenon in basketball shooting performance

Gabriel Calvo[1], *Carmen Armero*[2], *Luigi Spezia*[3]

[1]gabriel.calvo@ uv.es, Department of Statistics and Operations Research, University of Valencia

[2]carmen.armero@ uv.es, Department of Statistics and Operations Research, University of Valencia

[3]luigi@ bioss.ac.uk, Biomathematics & Statistics Scotland, Craigiebuckler, Aberdeen

Sports data analytics is a relevant topic in the statistical literature which is increasing in recent years. In basketball, a player or team is said to have a hot hand if their performance during a match is better than expected in a concrete period of time. This phenomenon has generated a great deal of controversy with detractors claiming its non-existence while other authors indicate its evidence. In this work, we present a longitudinal Bayesian hidden Markov model (HMM) that analyses the hot hand phenomenon in consecutive basketball shots, each of which can be either missed or made. Two possible states (cold or hot) are considered in the hidden chains, and the probability of success in each shot is modelled in terms of both the corresponding hidden state and the distance to the basket for that particular throw. This model is applied to a real data set corresponding to the Miami Heat team in the season 2005-2006 of the National Basketball Association (NBA). We show this model can be a powerful tool to assess the 'streakiness' of a team, and it provides information about the general performance during the match. In addition, the Bayesian HMM allows computing the posterior probability of any type of streak.

Keywords: Bayesian statistics, Bernoulli trials, Latent variables.

AMS:

Estimation of injury patterns according to maturity status and timing in an elite football academy based on zero-inflated models

Lore Zumeta-Olaskoaga[1,2], *Xabier Monasterio*[3,4], *Jon Larruskain*[3], *Jose A. Lekue*[3,4], *Juan M. Santisteban*[3,4], *Gontzal Diaz-Beitia*[3,4], *Dae-Jin Lee*[1]

[1]lzumeta@bcamath.org, dlee@bcamath.org,
BCAM - Basque Center for Applied Mathematics,
[2] Departamento de Matemáticas, Universidad del País Vasco UPV/EHU,
[3]Servicios Médicos, Athletic Club,
[4]Departmento de Fisiología, Universidad del País Vasco UPV/EHU

Injury prevention is a crucial task in the day-to-day routine of any football medical service and vast efforts are directed to it. Here, the first step to design an injury prevention strategy is to establish the extent of the injury problem. In this regard, two commonly used measures, to give a complete picture of injury risk, are: injury incidence and injury burden. They are calculated as number of injuries per player exposure (e.g. 1000 hours of exposure, days, season) and number of days lost due to injury per player exposure, respectively. The first calculates the rate at which new injury occurs (likelihood), whereas the second assesses how severe an injury is (consequences).

The second step is to identify factors and mechanisms that take part in the injury occurrence, in which statistical inference comes into play and special caution should be taken. Injury incidence and burden have characteristic shapes, seen as count data: they usually are right skewed and have many zeros. In this work, we focus on zero-inflated modelling approaches applied to epidemiological studies of sports injury data and we compare them with other widely used models.

We illustrate the application of different modelling approaches by a real study of a two-decade cohort of 110 growth curves of male academy football players aimed at estimating overall and specific injury burdens according to maturity status and timing. In this setting we show that the zero-inflated negative binomial mixed model is the best fitted model over other count regression models and that it enables a convenient way to analyse these type of injury data taking into account the time of exposure and within-player correlation. In addition, the model leads to more refined data analysis as it provides overall covariate effects, as well as separate effect estimates for the injury probability and for the injury burden mean in the respective logistic and truncated Poisson components. We encourage researchers in sports injury science to explore the distributional assumptions of their injury data and evaluate the choice of modelling.

Keywords: Zero-inflated models, sports injury prevention.

Parallel session: Ageing

Chairperson:

Maria Durban, Universidad Carlos III de Madrid (Spain)

Biological age imputation using data depth

S. Cabras[1], *I. Cascos*[2], *B. D'Auria*[3], *M. Durbán*[4], *V. Guerrero*[5]

[1]stefano.cabras@uc3m,es, [2]ignacio.cascos@uc3m.es, [3]bernardo.dauria@uc3m.es,
[4]marialuz.durban@uc3m.es, [5]vanesa.guerrero@uc3m.es
Department of Statistics, Universidad Carlos III de Madrid

The *biological age* is an indicator of the functional condition of an individual's body. Unlike the chronological age, which simply measures the time from birth, the biological age is also affected by the individual's medical condition, life habits, socio-demographics variables, as well as biomarkers, including some genetical ones. The assessment and use of the biological age is capturing increasing attention from the medical community, pharmaceutical companies, and also from the insurance companies in order to price their life and health insurances.

For ease of interpretation, the biological age is commonly given in the same unit as the chronological one (years) and an individual whose biological age is x years can be considered as someone whose body, from a functional viewpoint, and habits resemble the ones of a standard individual who is x years old. As a consequence, statistical *imputation* techniques can be used to assess the biological age of an individual whose chronological age is assumed to be missing.

In multivariate Statistics, the term *data depth* refers to the degree of centrality of an observation with respect to a data cloud or a probability distribution. The most central observations, which are highly representative of the whole of the data cloud, assume the highest depth values, while depth decreases as we move away from the centre towards outwarding observations.

Our proposal is to assess the biological age of an individual as the chronological age that would make her selected records as deep as possible when compared with those of other individuals in her age group (similar chronological age). This methodology can handle the situation when there is missing data in some specific variable different from the chronological age and can be adapted to a regression framework when there are few individuals in some given age group.

Keywords: Biological age, Data depth, Imputation

AMS: 62H05, 62H12

1

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

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

pgenius@barcelonabeta.org, [1]Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation, Barcelona, Spain. [2]Center for Genomic Regulation, Barcelona, Spain.

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

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

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

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

Keywords: Compositional Data Analysis; Imaging Genetics; Multivariate analysis; Selbal algorithm.

The new longevity in an inclusive society

Jose Miguel Rodríguez-Pardo[1]

[1]jmrpdc@gmail.com, Escuela de Pensamiento-Mutualidad de la Abogacía

First world societies see longevity as a problem: failure of the pension system, collapse of health care system, etc. However, longevity is the greatest achivement of humanity. In this talk we will present the challenges of the new longevity paradigm, and how a particular institution: Fundación Mutualidad de la Abogacía is contributing to this important problem. We will present several projects where the issues raised by the *new longevity* are apprached from different perspectives: economy, health, law, demography, etc, and how reserachers in many areas can contribute to this important topic by developing reserach projects in partnership with this institution.

Keywords: longevity, aging, inclusivity

Parallel session: Regional IBS

Chairperson:

Malu Calle, Universitat de Vic (Spain)

1

Resampling-Based Inference for High-Dimensional Regression

Anna Vesely[1], *Jelle J. Goeman*[2], *Livio Finos*[3]

[1]anna.vesely@unipd.it, Department of Developmental Psychology and Socialization, University of Padova

[2]j.j.goeman@ lumc.nl, Department of Biomedical Data Sciences, Leiden University Medical Center

[3]livio.finos@ unipd.it, Department of Developmental Psychology and Socialization, University of Padova

In linear regression, interest usually lies in discovering relevant predictor variables and assessing statistical significance; however, many challenges arise in high-dimensional settings, where the number of variables is potentially much larger than the sample size. We investigate the problem of making inference in this framework considering, for each predictor variable, the null hypothesis that the corresponding coefficient is zero. We build on the Multisplit method of Meinshausen et al. (2009) that computes adjusted p-values for each predictor, as well as the sign-flipping test of Hemerik et al. (2020).

We propose a novel procedure for multiple testing in high-dimensional regression. First, we introduce an approach that constructs permutation test statistics for each individual hypothesis by means of repeated random splits of the data. In each random split, half of the observations is used to perform variable selection, and half to build test statistics for the selected variables. Secondly, we use these statistics to efficiently test any subset of hypotheses. We define an asymptotically exact test by aggregating the individual statistics by means of a suitable function; different combining functions are possible, including the maximum and weighted sums. As the procedure gives a test for any subset of hypotheses, it can be embedded into closed testing methods to obtain simultaneous confidence bounds for the proportion of true discoveries (TDP). This way, we can make confidence statements on the TDP of all subsets, valid even under post-hoc selection.

Finally, we derive a second, approximate procedure that maintains the same asymptotic properties, but requires less memory usage and shorter computation time, and so can be scaled up to higher dimensions. The resulting method is extremely flexible, allowing different variable selection procedures and several combining functions; these may be chosen according to the desired power properties.

Keywords: high-dimensional linear regression; multisplit; resampling-based inference.

Flexible Inference in Multiverse Analysis

Paolo Girardi[1], Gianmarco Altoè[1], Antonio Calcagnì[1], Massimiliano Pastore[1], Livio Finos[1]

[1] Department of Developmental Psychology and Socialisation, University of Padova

Abstract

Data processing of non-trivial datasets often involves choices among several reasonable options for excluding, transforming, coding data, and modeling them. This multiplicity of steps gives rise to a multiverse of reasonable models and, therefore, statistical results. Unfortunately, it is a common practice to report only one privileged single analysis, therefore depriving the reader of the taste of this multiplicity and making the interpretation too optimistic.

Together with other questionable research practices, this is one of the main reason of the dramatic Reproducibility Crisis (Yong, 2012) and lack of confidence in many fields from psychology to economics, from sociology to medicine and neuroscience.

Steegen et al (2016) firstly proposed the Multiverse Analysis, which is, roughly speaking, nothing but the idea of frankly reporting all the analyses performed, then allowing the readers to evaluate the "stability" of the results. The proposal certainly represents an evaluable step toward the honest science. Since then, the method has been largely developed and has grown in popularity.

Despite this, it remains relegated to a descriptive role if as a formal inferential approach is not adopted. Simonsohn et al. 2020 firstly proposed a valuable method to derive a permutation-based test in this framework. However, this methodology is restricted to the linear model and does not cover all possible pre-processing steps. Furthermore – in our opinion – a more formal approach to the problem will cast the problem in the right theoretical context. In this contribution we exploit the flip-score test (Hemerik et al, 2020) to develop a very general and flexible approach that account for these issues.

- Yong, E. (2012). Replication studies: Bad copy. Nature News, 485(7398), 298.
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. Nature Human Behaviour, 4(11), 1208-1214.
- Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. Perspectives on Psychological Science, 11(5), 702-712.
- Hemerik, J., Goeman, J. J., & Finos, L. (2020). Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(3), 841-864.

1

A Joint Model for Multiple Longitudinal Outcomes, Recurrent and Terminal Events using CF Patient Registry Data

Pedro Miranda Afonso[1,2], *Dimitris Rizopoulos*[1,2], *Anushka Palipana*[3,4], *John P. Clancy*[5], *Rhonda D. Szczesniak*[3,4], *Eleni-Rosalina Andrinopoulou*[1,2]

[1]p.mirandaafonso@erasmusmc.nl, Department of Biostatistics, Erasmus Medical Center, Netherlands

[2]Department of Epidemiology, Erasmus Medical Center, Netherlands

[3]Division of Biostatistics & Epidemiology, Cincinnati Children's Hospital Medical Center, United States

[4]Department of Mathematics, University of Cincinnati, United States

[5]Cystic Fibrosis Foundation, Bethesda, United States

Cystic fibrosis (CF) is an inherited disease primarily affecting the lungs and gastrointestinal tract. It is of clinical interest to investigate the association between the risk of pulmonary exacerbation (PE) with lung function and nutritional decline, as direct positive associations between lung function and nutritional status have been reported. Previous work has been limited to continuous longitudinal markers and time-to-first PE, thereby neglecting subsequent occurrences and other survival outcomes. This was mainly due to the unavailability of appropriate and robust statistical software. Our primary goal is to simultaneously investigate the association between the risk of PE, lung function decline (FEV1), evolution of the patient's growth and nutritional status (e.g., BMI), and the risk of lung transplant or death using all available U.S. CF Foundation (CFF) patient registry data. We explore different forms of association between the longitudinal markers and the events of interest.

We propose a joint modeling framework accommodating multiple longitudinal markers, a recurrent event process, and a terminal event. The terminal outcome accounts for informative censoring due to lung transplantation or death from respiratory failure. Novel elements of our approach, compared to previously proposed joint models for recurrent events, are: (i) allowance for multiple longitudinal markers with different distributions, (ii) specifying various functional forms to link these markers with the risk of a recurrent event and the risk of the terminating event, and (iii) accommodation of discontinuous intervals of risk, and the time can be defined in terms of the gap or calendar timescale. The developed model is available in the R statistical package JMbayes2.

Analysis of all recurrent events with multiple biomarkers enhances our understanding of risks posed by PEs. Full MCMC algorithm implementation in C++ enables model fit in a timely fashion, despite its complexity. The proposed multivariate joint model affords the opportunity to make more efficient use of all available CFF registry data. It thereby brings new insights into CF disease progression and contributes to monitoring and treatment strategies.

Keywords: joint model, recurrent events, multivariate longitudinal data.

An adaptive design to handle deviations from proportionality assumption

Dimitris Karlis[1],*Urania Dafni*[2,3], *Panagiota Zygoura*[1],

[1]karlis@aueb.gr, Department of Statistics, Athens University of Economics and Business
[2]School of Nursing, National and Kapodistrian University of Athens
[3]Frontier Science Foundation Hellas

Most of the existing clinical trials designs for time to event outcomes are based on the proportionality of hazards assumption. Such an assumption implies that hazard ratio is a good summary and we can base the design upon it. There are a lot of example, as in the current cancer immunotherapy clinical trial designs, that creating the design on the assumption of proportional hazards can be an inapprorpiate assumption and, hence, lead to power loss or problematic inference at the end of the trial. In this work we try to overcome this problem by proposing an adaptive design.

Namely, we propose an adaptive design which, based on some interim look on the data, the assumption of proportionality is tested and then the design is adapted accordingly in order to achieve the required power and properties. The sample size can be updated based on the results of this interim look if the initial planned number found to be inaccurate.

We investigate the usage of different tests for proportionality assumption as well as different tests for updating the characteristics of the design in case of non-proportional hazards including trial design approaches based on genelarized log rank test and/or the restricted mean survival time. The properties of the adaptive design are investigated.

Keywords: Clincal Trials, proportionality

An adaptive design to handle deviations from proportionality assumption

Dimitris Karlis[1],*Urania Dafni*[2,3], *Panagiota Zygoura*[1],

[1]karlis@aueb.gr, Department of Statistics, Athens University of Economics and Business
[2]School of Nursing, National and Kapodistrian University of Athens
[3]Frontier Science Foundation Hellas

Most of the existing clinical trials designs for time to event outcomes are based on the proportionality of hazards assumption. Such an assumption implies that hazard ratio is a good summary and we can base the design upon it. There are a lot of example, as in the current cancer immunotherapy clinical trial designs, that creating the design on the assumption of proportional hazards can be an inapprorpiate assumption and, hence, lead to power loss or problematic inference at the end of the trial. In this work we try to overcome this problem by proposing an adaptive design.

Namely, we propose an adaptive design which, based on some interim look on the data, the assumption of proportionality is tested and then the design is adapted accordingly in order to achieve the required power and properties. The sample size can be updated based on the results of this interim look if the initial planned number found to be inaccurate.

We investigate the usage of different tests for proportionality assumption as well as different tests for updating the characteristics of the design in case of non-proportional hazards including trial design approaches based on genelarized log rank test and/or the restricted mean survival time. The properties of the adaptive design are investigated.

Keywords: Clincal Trials, proportionality

Parallel session: Bayesian analysis in medical studies

Chairperson:

Antonio López-Quilez, Universitat de Valencia (Spain)

Normative data of verbal fluency test in pediatric population from Spain: A Bayesian Regression Approach

Anabel Forte Deltell[1], *Laiene Olabarrieta-Landa*[2], *Juan Carlos Arango-Lasprilla*[3,4] *& Diego Rivera*[2]

[1]University of Valencia

[2] Public University of Navarre, Pamplona, Spain

[3] Biocruces Bizkaia Health Research Institute. Cruces University Hospital. Barakaldo, Spain

[4] IKERBASQUE. Basque Foundation for Science, Bilbao, Spain

The number of evoked words beginning with the phonemes m/p/r (phonological fluency) that children are able to say, turns out to be a good indicator of their performance of executive functions. Understanding how this value changes according to age, sex, and years of education of the families in charge, is important to make a correct diagnosis of possible learning problems.

For this study, 992 healthy boys (48%) and girls (52%) between 6 and 17 years old with an IQ ≥ 80 in the Nonverbal Intelligence Test and a value less than 19 in the Childhood Depression Inventory were evaluated. This is one of the largest verbal fluency studies carried out in Spain (Alicante, n=207; Almería, n=166; Granada, n=211; Madrid, n=171; Sevilla, n=167). Its results regarding the word count for each of the three phonemes have been studied using Bayesian Poisson regression models. In particular, the explanatory variables included in these models were selected using a Bayesian variable selection procedure.

The results of the study have been implemented in a web app that allows to compute the percentile for the number of evoked words (one per phoneme) of a given child adjusted to his/her demographic conditions, thus improving the diagnostic capacity of the practitioners.

Keywords: Bayesian models, Percentiles, Variable Selection

AMS: AMS Classification (Optional).

Bayesian modelling of the γ-H2AX assay for radiaton biodosimetry

Dorota Młynarczyk[1], *Pedro Puig*[2], *Carmen Armero*[3], *Virgilio Gómez-Rubio*[4]

[1]dorotaanna.mlynarczyk@uab.cat, Universitat Autònoma de Barcelona
[2]ppuig@mat.uab.cat, Centre de Recerca Matemàtica y Universitat Autònoma de Barcelona
[3]carmen.armero@uv.es, Universitat de València,
[4]virgilio.gomez@uclm.es, Universidad de Castilla-La Mancha

Biological dosimetry is a scientific discipline in which a variety of methods are used to estimate the radiation dose received by an individual exposed to accidental or therapeutic radiation. In real-life situations, timely determination of the absorbed dose is crucial to predict the health consequences of the irradiated patient and to ensure the appropriate treatment is selected quickly. A phosphorylated protein γ-H2AX is a well-established biomarker of ionising radiation, which is used to quantify the extent of damage at the cellular level. The main objective of this study is the development of new Bayesian methods which could be applied to γ-H2AX assay to address the different types of uncertainties arising during the dose estimation process.

The number of γ-H2AX foci is highly dependent not only on the absorbed dose but also on the time elapsed since exposure. For this reason, we present a new method in which a three-dimensional response surface is considered, i.e. the number of foci are both dose and time dependent. The Bayesian framework allows to incorporate the uncertainty about the moment of exposure and thus obtain more reliable results compared to the approach when the estimation is made at fixed post-irradiation times. Given the presence of overdispersion in γ-H2AX data, we used a finite Poisson mixture model for the distribution of foci counts, which is a meaningful alternative for such heterogeneous data.

Prior knowledge is frequently available in biological dosimetry, considering that in a real accident, some information about the potential dose and time will be provided. Laboratory experiments performed under controlled conditions can provide some reference data for the frequency of γ-H2AX foci and the level of radiation dose as well as the time elapsed since exposure. These results, in the form of calibration surface, may be then incorporated as a priori information into the model in order to estimate the dose for a newly exposed individual. This proposal allows for a significant reduction of the time needed to obtain results and the analysis suggests that such approach may be convenient in the dose estimation process using the γ-H2AX biomarker.

References

[1] López JS, Pujol-Canadell M, Puig P, Ribas M, Carrasco P, Armengol G, Barquinero JF. (2022). Establishment and validation of surface model for biodosimetry based on γ-H2AX foci detection. *International Journal of Radiation Biology*, 98(1), 1-10.

Keywords: bayesian inference, biodosimetry, gamma-H2AX

Bayesian hierarchical models for assessing the attitudes towards intimate partner violence against women

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

[1]antonio.lopez@uv.es, Depto. de Estadística e I.O., Universitat de València
[2]Depto. de Psicología Social, Universitat de València

This study is part of the research project that aims to map social attitudes towards intimate partner violence against women (IPVAW) in the city of Valencia in order to understand its determinants and analyse its influence on the incidence of this type of violence. The ultimate goal of this project is to determine whether the unequal geographical distribution of attitudes towards IPVAW influences the unequal distribution of the risk of IPVAW in neighbourhoods. For this purpose we have conducted an ecological study in small areas. For each of the 552 census tracts in the city of Valencia, information on attitudes towards IPVAW was obtained through a survey that integrates brief measures of five types of attitudes: perception of the seriousness of IPVAW, acceptability of IPVAW, victim-blaming attitudes, attitudes towards intervention in cases of IPVAW, and sexist attitudes. This survey integrates other individual-level data such as socio-demographic information and the perception of gender-based violence as a social problem, cohesion and informal social control in the neighbourhood. The study includes administrative information available on an aggregated basis for each of the census tracts.

A hierarchical model enables to explain the scores for each type of attitude in terms of personal characteristics and socioeconomic neighbourhood variables, incorporating interpersonal variability within the census tract and spatial variability. within the census tract and spatial variability. The maps obtained provide relevant pictures of attitudes, of the influencing terms, and of the homogeneity of neighbourhoods, as well as overall geographic patterns.

Attitude assessments are incorporated into the analysis of the risk smoothing model based on the protection orders recorded in the census tracts. Attitudes related to confirmed cases of IPVAW offer an improvement in the explanation of this social problem.

Keywords: Attitude mapping, Spatial smoothing, Violence risk

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

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

[1] Unidad de Bioinformática y Bioestadística. Insituto de Investigación Santiaria (INCLIVA)

[2] Digestive Disease Department, Hospital Clínico Universitario de Valencia, Spain.

[3] Institute of Liver Studies, Dept of Inflammation Biology, School of Immunology and Microbial Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom.

[4] Liver Failure Group, Institute for Liver and Disease Health, University College London, Royal Free Campus.

[5] European Foundation for the Study of Chronic Liver Failure (EF Clif) and the European Association for the Study of the Liver-Chronic Liver Failure (EASL-CLIF) Consortium

Corresponding author: jacarbonell@incliva.es

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

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

With our preliminary results, the risk of suffering ACLF increases the hazard by a factor of 1.150 by each unit increase of A-ULN, with a 95% credible interval that ranges between a lower limit of 0.93 and an upper limit of 1.405. The probability of the coefficient for β to be greater than 0, and therefore considering A-ULN a predictor for ACLF is 0.91. However, the Weibull distribution used to specify the baseline hazard function could not adjust the variability of the 3 hospitals included, forcing us to reduce the number down to 2. We are planning on testing a mixture of piecewise constant functions to adjust the baseline hazard function that could allowed us to incorporate data from all three hospitals to the final model. Moreover, we intend to include hospital as a random effect factor using a competing risk frailty model, which could also possibly solve the baseline hazard function convergence problem as well. Finally, we expect to generate a multivariable competing risk frailty model in the following months using this data, including again hospital id as a random effect factor.

Keywords: Survival Analysis, Bayesian, Competing Risks.

Bayesian adaptive design for a clinical trial on cardiology

Stefano Cabras[1], *María Eugenia Castellanos*[2], *Alicia Quirós*[3]

[1]stefano.cabras@uc3m.es, Department of Statistics, Universidad Carlos III de Madrid
[2]maria.castellanos@urjc.es, Department of Computer Science and Statistics, Universidad Rey Juan Carlos
[3]alicia.quiros@unileon.es, Department of Mathematics, Universidad de León

The aim of this work was to calculate the sample size and define an adaptive design for a randomised study to demonstrate the superiority of one device for detecting atrial fibrillation (AF) compared to another device. Patients enter the study when implanted with one of the two devices, randomly assigned, and will be followed up for 24 months to observe the detection of AF de novo by the device. The study's main hypothesis assesses whether $H_1 : p_A > p_B$, with A being the new device group and B being the comparison device group.

As we propose an adaptive design, the Bayesian approach arises naturally. The estimation of the proportions will be done with a Beta-Binomial model and the comparison with the Bayes factor (BF). Success in the study will be obtained if $BF > 10$, which is equivalent to accept H_1 if $P(H_1|\text{data}) > 0.909$. Both design and weak informative analysis prior distributions are defined from previous studies in the form of power priors (Ibrahim and Chen, Statist Sci, 2000).

Sample size has been estimated following Wang and Gelfand (Statist Sci, 2002). Simulations of different sample sizes for each group have been carried out, using the prior predictive distribution from the design priors. In each simulation, the study's success was determined and we chose the minimum sample size such that power is at least 80%.

An interim analysis is established 27 months after the first patient inclusion. At that time, two criteria are used to assess whether to stop recruiting patients, either because study failure is anticipated or study success is expected to be achieved (Berry et al., CRC Press, 2011) without the need to observe more patients.

Keywords: Adaptive design, Bayes factors, Sample size determination

Parallel session: Medical studies 2

Chairperson:

Christian Tebé, IDIBELL, Unitat de Bioestadística (Spain)

1

A comparative study of quality of life among patients with metastatic breast cancer and breast cancer survivors: The chronic effect of the disease

Francisca Morey [1,3], Eva Rodríguez[2], Lorena Gómez Villarroya[2], Clara Serra-Arumi[2,5], Coral Báez Sáez[2], Miguel Gil-Gil [6,7], Gloria Campos[6], Andrea Vethencourt[6], Silvia Vázquez[6], Agostina Stradella[6], Sonia Pernas [6,7], Antoni Font[2], Catalina Falo [6,7], Marisa Mena [1,3,4$], Sara Tous [1,3,4$]

$Both authors contributed equally as co-senior authors

1. Cancer Epidemiology Research Program (CERP), Institut Català d'Oncologia. L'Hospitalet de Llobregat, Barcelona.
2. Facultad de Psicología, Universitat Autònoma de Barcelona. Barcelona.
3. Epidemiology, Public Health, Cancer Prevention and Palliative Care Program. Institut d'Investigació Biomèdica de Bellvitge (IDIBELL). L'Hospitalet de Llobregat. Barcelona.
4. CIBER en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain.
5. Teaching, Research and Innovation Unit. Institut de Recerca Sant Joan de Déu. Esplugues de Llobregat. Barcelona.
6. Medical Oncology department. Institut Català d'Oncologia. L'Hospitalet de Llobregat, Barcelona.
7. Breast Cancer Group, Institut d'Investigació Biomèdica de Bellvitge - IDIBELL, L'Hospitalet de Llobregat, (Barcelona) Spain.

Keywords: breast cancer (BC), quality of life (QoL), clustering.

2

Breast cancer (BC) is the most common malignancy among women with an annual incidence of approximately two million new cases worldwide. The five-year overall survival has improved in the last 15 years (83%). However, about 20% of all BC patients deal with a diagnosis of metastatic disease. Alongside the increased quantity of life (QoL), however, quality of life issues arise. The aim of this investigation is to evaluate the chronic effect of BC by assessing the QoL of two groups of patients, one with metastatic breast cancer (MBC) and the second with early-stage breast cancer (EBC).

QoL was assessed with Afex-CA-QL questionnaire. Clinical and sociodemographic characteristics were collected from medical records. Descriptive analyses were performed to explore the differences between the MBC and EBC patients. And then cluster analysis was used to classify the patients according to their clinical, sociodemographic and QoL characteristics. To determine the differences in QoL between patients, the distance matrix between the variables was computed with the Gower distance. The analysis was performed in two-steps: in the first step, a Hierarchical Cluster (HC) was computed in order to determine the optimal number of clusters (k), calculated with an agglomerative strategy (bottom-up approach): each observation starts in its own cluster, and pairs of clusters are merged resulting in a new single cluster at higher level. The results of the HC are presented in a dendrogram (tree structure). The second step, the k-means algorithm, was used to distribute each observation into the k optimal clusters with the nearest mean. The result is presented in a multidimensional scaling analysis with a two-dimensional plot. Finally, the cluster's profiles were compared regarding the sociodemographic characteristics of the patients included in a bivariate analysis.

Finally, 50 patients with MBC and 160 patients with EBC were included in the analyses. Aside clinical characteristics, MBC differed from EBC patients in working situation and psychopharmacological treatment adherence/use, but not in QoL. Cluster analysis identified three groups of patients (figure 1): two with high QoL, first including mostly MBC patients and the second including mostly EBC patients, and the third, with low QoL, including both MBC and EBC patients. The cluster's profiles showed that patients from the group with low QoL were significantly younger and more unemployed than patients from the other two groups (p = 0.038).

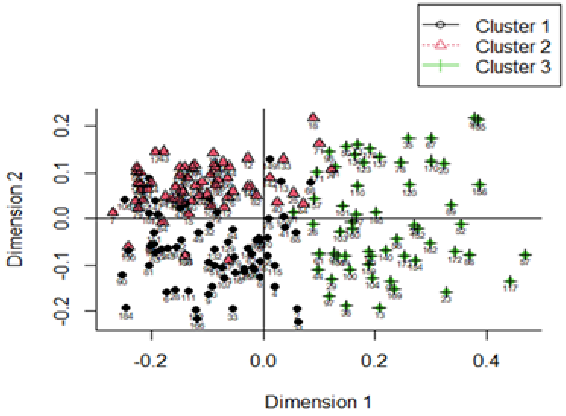


Figure 1 Two dimensional plot resulting from cluster analysis

In conclusion, our results indicate that BC patients have a decrease in certain QoL aspects but did not present QoL differences by disease status, as was the initial hypothesis, but for working status and age. The cluster analysis allowed to classifying patients from a completely unbiased point of view, as by definition all model variables were introduced as independent and it is the algorithm that calculated how similar they are, generating the different clusters and determining that QoL was an aspect that discriminated against clusters. These results must be taken into account to further optimize care for this growing population. This knowledge may help health care providers, policymakers, and researchers to appropriately and holistically respond to the diverse care and support needs of those living with BC, and will help to promote optimal QoL for all women.

Generative adversarial networks for medical images editing

Esteban Vegas[1], *Ferran Reverter*[1], *Rubén Fernández*[1]

[1]evegas@ub.edu, freverter@ub.edu, rubenfb14@gmail.com Department of Genetics,
Microbiology and Statistics, University of Barcelona

Medical imaging applications based on Deep Learning models have been growing in the last decade and it seems that the trend will continue in the coming years. Segmentation, pattern detection or classification are some of the most common tasks performed by these type of models. However, it has a major drawback, the need of large amounts of tagged data. This fact has an even greater impact in this specific field, as in contrast with more general areas, labeling must be carried out by experts. This leads to a more important shortage on labeled images. The importance of unsupervised methods such as some types of Generative Adversarial Networks (GANs), that can operate without any prior labeling is, for the aforementioned reasons, justified.

GANs are new architectures (2015) that use two neural networks, called generator and discriminator, in order to generate new, synthetic instances of data that can pass for real data. In images analysis, the generator learns to generate plausible images. The generated instances become negative training examples for the discriminator. The discriminator learns to distinguish the generator's fake images from real ones. Besides, the generative model in the GAN learns to map points in the latent space to generated images. This latent space has structure that enable the semantic editing of images. This type of editing allows the modification of high-level features of the images by performing vector arithmetic in the low-dimensional latent space learned by the GAN.

This work considers a set of arithmetic operations in the latent space of GANs for editing histopathological images. We analyze thousands of image patches from whole-slide images of breast cancer metastases in histological lymph node sections. Image files were downloaded from the pathology contests CAMELYON 16 and 17. We show that widely known architectures, such as: Deep Convolutional Generative Adversarial Networks and Conditional Deep Convolutional Generative Adversarial Networks, allow image editing using semantic concepts that represent underlying visual patterns in histopathological images, expanding GAN's well-known capabilities in medical image editing.

Keywords: Deep learning, Generative adversarial networks, Medical image editing

AMS: 68T07, 68T09

Predicting multidrug resistance in neutropenic cancer patients with bloodstream infection due to *Pseudomonas aeruginosa*

Pallarès N[1]*, Satorra P*[1]*, Albasanz-Puig A*[2]*, Laporte-Amargós J*[2]*, Gudiol C*[2]*, Tebé C*[1]

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

[2]aalbasanz@bellvitgehospital.cat, jlaporte@iconcologia.net, cgudiol_ext@iconcologia.net. Infectious Diseases Department, Bellvitge University Hospital

Background: Infections caused by multidrug-resistant bacteria have become a therapeutic challenge in patients with febrile neutropenia, which is a very common complication in onco-hematological patients receiving chemotherapy, and which is associated with high morbidity and mortality.

Objective: To develop a clinical prediction model to estimate multidrug resistance risk in onco-hematological patients with neutropenia and bloodstream infection (BSI) due to multidrug-resistant *Pseudomonas aeruginosa* (MDRPA).

Methods: A multicenter, retrospective cohort study including neutropenic patients with BSI due to MDRPA was conducted in 34 centers (12 countries) between 2006 and 2018. The study sample was split into a derivation (80%) and a validation cohort (20%).

A mixed logistic regression model with random intercept (center) was used to develop a predictive model. Ten multiple imputations with chained equations (MICE) were used to deal with missing data. Each of the 10 datasets created was sampled by bootstrapping with replacement 100 times, totaling 1000 samples. A model was fitted in each sample using backward elimination (AIC criteria). Predictors retained in more than 80% of the 1000 estimated models were considered for inclusion in the final model. Rubin's rule was used to summarize a model with the selected predictors. Discrimination was assessed by estimating the area under de ROC curve (AUC), calibration by comparing observed versus expected MDR by tenths of predicted risk. All validation analyses performed in the derived sample was also repeated in a preserved sample for validation.

Results: Among 1217 episodes of BSI (38.3% women, 57.8-yo, 75.3% hematologic disease), 25% were caused by an MDR strain (95%CI 19.7 to 30.9). After bootstrapping process elected variables were prior therapy with piperacillin/tazobactam, prior antipseudomonal carbapenem use, fluoroquinolone prophylaxis, hematological underlying disease, presence of a urinary catheter, and age. The predictive model obtained in the derivation cohort had excellent discrimination with an AUC of 0.82 (95% CI 0.79-0.85). The observed probability corresponded well to the predicted probability. The performance in the derivation and validation cohort was similar. A Shiny app (https://ubidi.shinyapps.io/ironic/) was developed to calculate the risk of MDR.

Conclusions: This prediction model will improve the identification of patients at high risk for multidrug resistance and will avoid the use of broad-spectrum antibiotics in patients with less risk of resistance.

Keywords: Infectious diseases, Prediction model, Bootstrapping, Shinny app.

The importance of data visualization for a better interpretation of the results: application to a multinomial analysis of the positivity of different biomarkers on the risk of oropharyngeal cancer

Sara Tous[1], *Francisca Morey*[2], *Marisa Mena*[2], *Laia Alemany*[1], *Miren Taberna*[3,4], *Hisham Mehanna*[5] *on behalf of the EPIC-OPC group*

[1]Cancer Epidemiology Research Program, Institut Català d'Oncologia (ICO)-IDIBELL, CIBERESP, L'Hospitalet de Llobregat, Barcelona, Spain

[2]Cancer Epidemiology Research Program, IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain

[3]Medical Oncology Department, Head and Neck Cancer Unit, Institut Català d'Oncologia (ICO), L'Hospitalet de Llobregat, Barcelona, Spain

[4]Scientific Officer, Savana, Madrid, Spain

[5]Institute of Cancer and Genomic Sciences Robert Aitken Building, The University of Birmingham, Institute for Cancer Studies, Birmingham, Great Britain

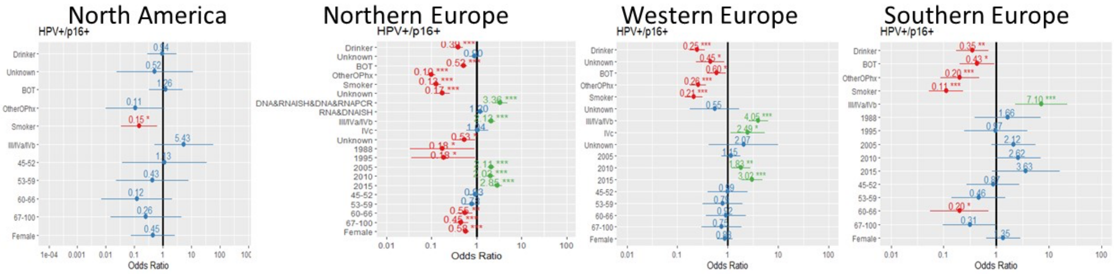
Keywords: Human Papillomavirus (HPV), p16 Immunohistochemistry, Oropharyngeal Cancer (OPC), Multinomial regression, Data visualization.

AMS: AMS Classification (Optional).

Generalized Linear Model (GLM) unifies the linear, logistic and Poisson regression models and allow the response variables having a different distribution of errors than normal. The logistic regression, using as link function the logit, is used to model a binary dependent variable. Multinomial regression model is an extension of the logistic regression model with the particularity that the outcome (the variable we want to explore) has more than two categories. This model allows us to explore the relationship of a categorical variable with a set of independent variables to predict events. From the coefficients of the model, we can interpret the effects these variables have on the response. Let Y be a categorical variable with J categories and y its $\pi_1, \pi_2, \ldots, \pi_J$ related probabilities, being $\sum_{j=1}^{J} \pi_j = 1$. The multinomial regression model is constructed considering one category as the baseline response, the last one J, for instance, and is defined as a logit model related to this reference as follows:

$$ln(\pi_j/\pi_J) = \alpha_j + \sum_{k=1}^{K} \beta_{jk}X_{jk},\ j = 1, \ldots, J\text{-}1.$$

The model has J-1 equations with its own parameters. Each of these parameters express the effect respect to the reference category. The parameters are estimated using the maximum likelihood method and we can check if a variable is significant using Wald's statistic and the conditional contrast of the reason of verisimilitude. The efficiency of the model is analysed using the likelihood ratio. In the context of an international study comprising 13 cohorts of patients with oropharyngeal cancer (OPC) we collected data on p16$^{\text{INK4a}}$ (p16) expression (+ if overexpressed and – if under expressed the protein), Human Papillomavirus (HPV) detection (+ if positive and – if negative), demographics, tobacco/alcohol use and clinical data. A centralized individual patient data reanalysis was performed. A fraction of OPC cases are caused by HPV and p16 is considerate a surrogate marker of HPV oncogenic activity. We generated a combination variable joining p16 and HPV results obtaining a variable of study that had four categories: p16-/HPV- (double negative), p16+/HPV-, p16-/HPV+, and p16+/HPV+ (double positive). We were interested on clearly define the proportion and determinants of OPC patients who were p16+ but HPV-, globally and by geographical regions. Hence, we adjusted a multinomial regression model to evaluate the factors associated with the different biomarkers combinations compared to the double negative cases, globally and by geographical regions. In total, 7654 patients diagnosed between 1988 and 2018 from nine different countries were included. Regionally, 186 cases were from North America, 3964 from Northern Europe, 2647 from Western Europe and 857 from Southern Europe. Regarding p16 and HPV detection, 3560 cases were double negative, 415 p16+/HPV-, 289 p16-/HPV+ and 3390 double positive. Among p16+ cases, 10.9% were HPV-. This proportion differed significantly by cohorts and geographic regions (p-value<0.001) and was lowest in the highest HPV prevalence areas. The determinants of p16+/HPV- cases also differed by regions. The huge tables produced as the result of the multinomial analysis led us to present graphically the results to its better interpretation. We generated homogeneous plots for each region and p16/HPV combination that helped us to compare the results obtained.



Colours legend: protective factor (in red), risk factor (in green), no significant association (in blue).

Regression analysis with interval-censored covariates. Application to liquid chromatography.

Klaus Langohr[1], *María Marhuenda Muñoz*[2], *Guadalupe Gómez Melis*[1]

[1]klaus.langohr@upc.edu, lupe.gomez@upc.edu, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya · BarcelonaTech, Barcelona
[2]mmarhuendam@ub.edu, Department of Nutrition, Food Sciences and Gastronomy, Universitat de Barcelona

Interval-censored observations of a response variable are a common occurrence in medical studies and usually result when the response is the elapsed time until some event whose occurrence is periodically monitored. A different situation is encountered when the explanatory variables are interval-censored and the response variable is completely observed.

In this work, we apply and extend the EM-type algorithm (GEL algorithm) developed by Gómez, Espinal, and Lagakos (2003) to fit a linear regression model with an interval-censored covariate to a novel approach in metabolomics. Compounds, such as plasma carotenoids, are quantified by liquid chromatography coupled to mass spectrometry and have detection and quantification limits because of their low quantity in the body and the quantitation methods' limitations; as a result, the compound concentrations are interval-censored. The sums of some of these compounds are also of great interest since they help to understand the global action and possible synergy of a family of compounds. These sums are also interval-censored random variables with lower and upper limits given by the sums of the lower and upper limits of all the compounds, respectively.

The methodological novelty presented in this work is the extension of the GEL algorithm to generalized linear models. We present the fit of gamma and logistic regression models with an interval-censored covariate and adapt the corresponding Pearson and deviance residuals to these specific models. Data analysis on the association of metabolites extracted from human samples and measured by liquid chromatography with anthropometric, clinical, and biochemical parameters is used as illustration.

So far, there is no consensus as of how data under the limit of quantification or even detection should be treated. In this context, our method works as a reliable way of using the real known (or unknown) values in order to interpret their role in health and disease.

Keywords: Interval-censored covariates; Generalized linear model; Liquid chromatography.

Second session young researchers

Chairperson:

Ana Justel, Universidad Autonoma de Madrid (Spain)

Operating characteristics of a model-based approach to incorporate non-concurrent controls in platform trials

Pavla Krotka*, Martin Posch, Marta Bofill Roig

Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria

*pavla.krotka@meduniwien.ac.at

Abstract

To evaluate the efficacy of multiple treatments compared to a shared control in a platform trial, we consider a model-based approach to use non-concurrent control data. As not all treatments are studied at the same time, bias might occur due to possible time trends. The model-based analysis adjusts for the factor time to avoid bias. We examine the proposed method under a range of scenarios and outline under which conditions it provides valid inference.

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

1. Introduction

Platform trials are multi-arm multi-stage clinical trials that aim at evaluating the efficacy of several treatment arms within a single trial. Treatments are permitted to enter the trial as they become available, and leave at different times, so the total number of treatments under study is not pre-specified. Moreover, treatment arms may be compared against a shared control group. For treatments that enter the trial later, the control group is divided into concurrent (CC) and non-concurrent controls (NCC), i.e., patients randomized before the given treatment arm entered the trial. While using NCC data can offer several benefits, such as reduced sample size and increased statistical power, it may also introduce bias in effect estimators if time trends exist [1].

Modeling approaches to include NCC have been discussed in platform trials where one treatment arm is added later [2, 3]. In this work, we consider a regression model for treatment-control comparisons that allows for the late addition of several treatments and incorporation of NCC while adjusting for time trends and investigate through simulations when it leads to valid inference. Especially, we consider trials with continuous endpoints and linear and stepwise time trends. We investigate the performance of the proposed model in terms of the type I error rate and power of the individual treatment-control comparisons and examine the behaviour of the power with respect to the entry time of treatment arms. We show by simulations that type 1 error control is lost if the assumption of equal time trends in all arms is violated. Furthermore, we quantify the power gain by using NCC compared to tests using only CC or tests naively pooling CC and NCC.

2. Regression model with time trend adjustment for incorporating NCC

Consider a platform trial evaluating the efficacy of K treatments compared to a shared control. Assume that treatment arms enter the platform trial sequentially. We divide the duration of the trial into S periods, where the periods are the time intervals bounded by any treatment arm either entering or leaving the platform (see Figure 1 A for illustration). Suppose that patients are randomised between control and active treatment arms using 1:1:...:1 allocation ratio in each period. We denote by k the treatment group where $k = 0$ indicates the control and $k = 1, \ldots, K$ the treatment arms, ordered by entry time. We assume that the sample size n of arms $k = 1, \ldots, K$ is equal across all treatment arms. Besides, we assume that the trial starts with one treatment arm and the control and the remaining treatments enter in a staggered way. Especially, we consider a trial where a new arm enters after every δ patients have been enrolled in the control. Thus, the total sample size is given by $N = (K - 1)\delta + (K + 1)n$. Note that if $\delta = 0$, all arms enter the trial simultaneously, while if $\delta = n$, a new arm comes in when the previous arm is completed. The observed data can be summarized by $\mathcal{D} = \{(y_j, k_j, s_j), j = 1, ..., N\}$, where j represents the patient index, y_j denotes the continuous response and k_j and s_j are the arm and period corresponding to a given patient j.

We focus on the inference for arms $k = 2, 3, \ldots, K$ that enter when the trial is already ongoing and therefore NCC data is available. We aim at comparing the efficacy of each treatment against the control as soon as the treatment arm is completed. Consider the test of the null hypothesis $H_{0M} : \theta_M \leq 0$ for arm M under study, where θ_M denotes the treatment effect size for treatment M. To test H_{0M}, we propose a regression model fitted using all data from the trial until treatment M leaves the platform (i.e., all data in the set $\mathcal{D}_{A(M)} = \{(y_j, k_j, s_j), j = 1, ..., N; s_j \leq S_M\}$, where S_M denotes the period in which arm M finishes). The model estimates the effects of the treatments in the trial up to period S_M, adjusts for time using a step function, and is given by:

$$E(y_j) = \eta_0 + \sum_{k \in \mathcal{K}_{A(M)}} \theta_k \cdot I(k_j = k) + \sum_{s=2}^{S_M} \tau_s \cdot I(s_j = s) \qquad (1)$$

where η_0 is the response in the control arm in the first period, and τ_s denotes a stepwise period effect between periods 1 and s. θ_k represents the effect of treatment k compared to control, for $k \in \mathcal{K}_{A(M)}$, where $\mathcal{K}_{A(M)}$ is the set of treatments that were active in the trial during periods prior or up to S_M. In this model, data from all arms that entered the platform so far contribute to the estimator of the period effect τ_s. Note that this model assumes a time trend that is constant in all periods and equal across all treatment arms. We evaluate the performance of (1) in scenarios where the model assumptions are met, as well as in cases where they are violated and time trends differ across arms or are not constant over periods. For comparative purposes, we also analyze the data with two standard approaches, where the data from the investigated treatment arm M is compared either to pooled NCC and CC data (pooled analysis) or only CC data (separate analysis) with one-sided t-tests.

3. Assessment of analysis approaches through simulations

3.1. Data generation and simulation design

We evaluated the performance of the methods through simulations. We simulated trials with three treatment arms ($K = 3$) as described in Sect. 2, where new arms enter after every

$\delta \in \{0, 30, 50, 70, 90, 100, 130, 150, 167\}$ patients were recruited in the control arm. To compare the different scenarios, we fixed the overall sample size $N = 1000$. Given δ we then chose the sample size per treatment n, such that the overall sample size is N. We assumed that the patient index corresponds to the order in which patients are enrolled and that at each unit of time exactly one patient is recruited so that $t = j$ for $j = 1, ..., N$. Patients were assigned to arms following block randomization with block sizes of $2 \cdot (\#\text{active arms} + 1)$ in every period. The data was generated according to

$$E(y_j) = \eta_0 + \sum_{k=1,2,3} \theta_k \cdot I(k_j = k) + f(j), \tag{2}$$

where η_0 and θ_k are the response in the control arm and the effect of treatment k as in (1). The function $f(j)$ denotes the time trend, which is either linear – given by $f(j) = \lambda_{k_j} \frac{(j-1)}{(N-1)}$ – or stepwise – given by $f(j) = \lambda_{k_j}(s_j - 1)$ –, where λ_{k_j} quantifies the strength of the time trend. Note that, as k_j takes values in $\{0, 1, 2, 3\}$, λ_{k_j} can be different depending on the arm. Whenever $\lambda_{k_j} \neq \lambda, \forall j$, it implies different time trends with respect to arms. We assumed normally distributed residuals with variance $\sigma^2 = 1$ across treatment and control arms and zero mean for control arm ($\eta_0 = 0$). We simulated trials under the null hypothesis ($\theta_k = 0$, for $k = 2, 3$, while $\theta_1 = 0.25$), as well as under the alternative, using an effect of $\theta_k = 0.25$ for all treatment arms ($k = 1, 2, 3$). We considered linear and stepwise time trend patterns with equal time trends across all arms ($\lambda_k = \lambda, \forall k$), and cases with a different time trend in arm 1 but equal across the remaining arms ($\lambda_0 = \lambda_2 = \lambda_3 \neq \lambda_1$). We tested H_{0M} for $M = 2, 3$ at significance level $\alpha = 0.025$. We simulated 100,000 replicates for each scenario to estimate the type 1 error and statistical power.

3.2. Model performance in terms of type I error and power

For scenarios simulated under the null hypothesis, all investigated analysis approaches lead to an estimated type I error of 2.5% when there are no time trends. When time trends exist but are equal between arms, then the pooled analysis gives an inflation in the type 1 error. The proposed model maintains the type I error at 2.5%, even if the time trend is linear, and so does the separate analysis which only uses CC data. When there are different trends in the arms, the control of the type I error is lost when using the proposed model: Figures 1 C and D depict the type I error and power of H_{0M} for $M = 2, 3$ for fixed stepwise time trends $\lambda_k = 0.1$ in arms $k = 0, 2, 3$ and varying time trends in arm 1. Note that the inflation in type I error of the regression model is lower for arm 3 than for arm 2. This is because the impact of the violation of the equal time trends assumption is lower the more arms there are in the trial, as the contribution of the arm with a different time trend to the estimation of the time effect is smaller.

Under the alternative hypothesis and no or equal time trends, for scenarios in which treatment arms enter the trial in a staggered way ($\delta > 0$), there is a gain in power when using the proposed regression model as compared to the separate analysis, while the largest power is reached with the pooled analysis. Also, the power of the treatment-control comparisons in the platform trial increases for arms that were added later since the size of NCC is larger. Therefore, the power of testing treatment arm 3 is larger than the power for arm 2. Furthermore, we observe that the amount of time that arms overlap in the trial have also implications for the power (see Figure 1-B). For the regression model and the separate analysis, it is more efficient to evaluate the arms simultaneously. Specifically, the more treatments overlap, the more powerful the platform is. This is due to the fact that less sample size is required in the control when it is shared across treatment arms. On the other hand, the pooled approach may even give larger power for intermediate δ.

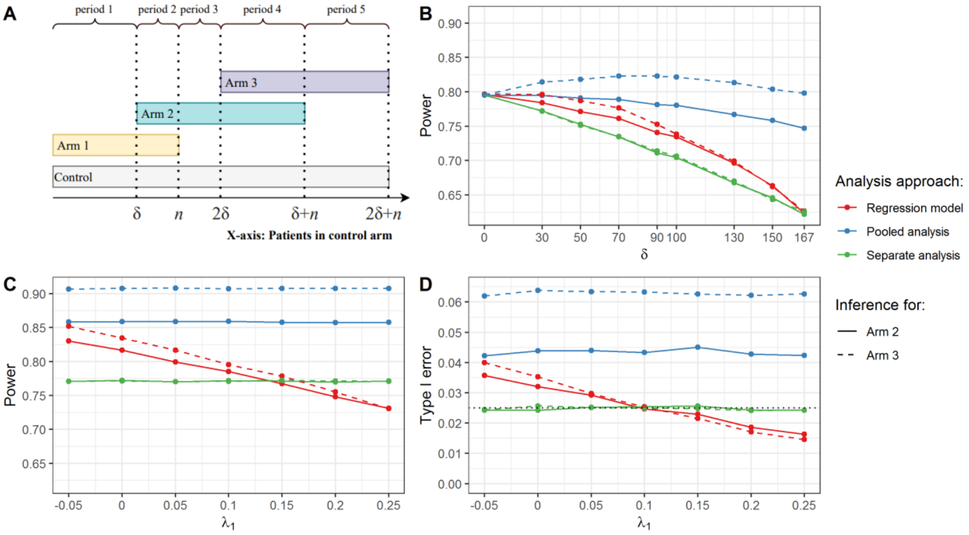


Figure 1: **A.** Platform trial scheme. **B.** Power for rejecting H_{0M} ($M = 2, 3$) in case of no time trends ($\lambda_k = 0, \forall k$) and varying arm entry times δ. **C-D.** Power and type I error for $\delta = 30$ and unequal stepwise time trends: $\lambda_k = 0.1$ for $k = 0, 2, 3$, and varying time trend λ_1 in arm 1.

4. Discussion and conclusions

We considered a regression model to incorporate NCC in platform trials and evaluated its operating characteristics in the presence of time trends. In the investigated scenarios where time trends are equal across all arms, the model controls the type I error and improves the power as compared to using only CC data for the analysis. This holds even if the model is misspecified and the time trend is linear rather than stepwise as fitted in the model. Although we focused on continuous endpoints in this work, the proposed method could be extended to binary and survival data by considering similar adjustments in the corresponding logistic and Cox models.

Platform trials provide an efficient framework to test multiple treatments and are more flexible than classical multi-arm trials, where the number of treatments under evaluation is pre-specified. However, the timing of adding arms affects the power: this decreases when arms are added at later times. Even so, platform trials can prove powerful compared to separate trials for individual treatments, as sharing controls reduces the required sample size. Incorporating NCC can further improve the power, but to achieve sound results, analyses should be adjusted for time trends and the plausibility of model assumptions (such as equal time trends) carefully assessed.

5. Acknowledgements

EU-PEARL (EU Patient-cEntric clinicAl tRial pLatforms) project has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking (JU) under grant agreement No 853966. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and Children's Tumor Foundation, Global Alliance for TB Drug Development non-profit organisation, Spring works Therapeutics Inc. This publication reflects the authors' views. Neither IMI nor the European Union, EFPIA, or any Associated Partners are responsible for any use that may be made of the information contained herein.

6. Bibliography

[1] Park, J.J.H., et al. (2022). *How to Use and Interpret the Results of a Platform Trial.* JAMA.

[2] Lee, K.M., et al. (2020). *Including non-concurrent control patients in the analysis of platform trials: is it worth it?*. BMC Med. Res. Methodol.

[3] Bofill Roig, M., et al. (2021). *On model-based time trend adjustments in platform trials with non-concurrent controls.* arXiv:2112.06574 (https://arxiv.org/abs/2112.06574).

Modeling COPD hospitalizations using variable domain functional regression

Pavel Hernández Amaro[1], *María Durbán Reguera*[2], *María del Carmen Aguilera Morillo*[3], *Cristobal Esteban Gonzalez*[4], *Inma Arostegui*[5]

[1]pahernan@est-econ.uc3m.es, Departmento de Estadística, Universidad Carlos III de Madrid
[2]mdurban@est-econ.uc3m.es, Departmento de Estadística, Universidad Carlos III de Madrid
[3]mdagumor@eio.upv.es, Universidad Politécnica de Valencia
[4]CRISTOBAL.ESTEBANGONZALEZ@osakidetza.eus, Osakidetza servicio vasco de salud
[5]inmaculada.arostegui@ehu.eus, Universidad del País Vasco

Abstract

In this work we present the use of variable domain functional regression models to estimate the number of hospitalizations from COPD patients and its relation to physical activity. The functional variable is the number of daily steps walked by each patient. We use a bi-dimensional basis representation of the functional coefficient and estimate the model via Penalized Quasi-Likelihood using the mixed model representation of a penalized spline.

Keywords: Variable domain functional regression, penalized spline, COPD.

1. Introduction

Variable domain functional regression models are an extension of scalar-on-function models, where the functional predictor is observed on a grid of different length for each subject. Such type of data have become very common recently, specially due to the wide-spread use of wearable devices and their ability to collect data that can improve, for example, health diagnostics.

We base our regression model in the one presented in [3], but we improve their approach by assuming that the functional covariate is smooth but observed with error. This assumption add further complexity to the estimation procedure, but the results of our simulation studies (not included due to lack of space) show a better performance of our approach.

Our data come form the TELEPOC study carried out at the Galdakao University Hospital (Bilbao, Spain) between the years 2010 and 2013, this study collects a wide range of data from 110 patients suffering Chronic Obstructive Pulmonary Disease(COPD), including smoking habits, sociodemographic characteristics such as age or gender, clinical data such as anxiety or depression symptomatology and the number of hospitalizations suffered by each patient due to CPOD during its time in the study [2]; the study also keeps track of the daily number of steps given by the patients.

The number of days where steps are collected is different from patient to patient. The are many reasons for this: some patients die during the study, not all patients enter the study at the same time and some patients simply abandon the study before it ends. Therefore we are in presence of variable domain functional data. We apply our methodology to estimate the relationship between physical activity, measured as daily steps, and the annual ratio of hospitalizations due COPD. This relationship is directly reflected by the functional coefficient of the model.

2. Methodology

We have the following sample data: $\{Y_i, Z_i, X_i(t)\}$, $i = 1, \dots, 110, t \in [1, T_i]$, where Z_i are the non-functional covariates, $X_i(t)$ is the functional covariate: daily steps walked by patient i at day t and T_i is the length of the domain of variable t for subject i, i.e. the number of days patient i participated in the study.

Finally our response variable Y_i is the number of hospitalizations due to COPD suffered by patient i during his/her time in the study, hence this variable follow a Poisson distribution, $Y \sim Poi(\mu)$. As can be expected, patients with more time in the study will have more hospitalizations, to avoid this cumulative effect we will use the annual ratio of hospitalizations as our response variable. Our variable domain functional regression model is then:

$$\eta = log(\mu) = \alpha + Z\gamma + \frac{1}{T}\int_T X(t)\beta(t,T)dt + log\left(\frac{T}{365}\right), \ \ t \in [0,T]. \tag{1}$$

where η is our linear predictor and our link function is $g(\mu) = log(\mu)$. To fit this model our first step is to make a basis representation of the functional variable $X(t)$ and the bi-dimensional functional coefficient $\beta(t,T)$:

$$X(t) = \sum_{j=1}^{p} a_j \phi_j(t) = \boldsymbol{\phi a}, \quad \beta(t,T) = \sum_{l=1}^{q}\sum_{k=1}^{r} b_{lk}\varphi_l(t)\psi_k(T) = (\boldsymbol{\varphi} \otimes \boldsymbol{\psi})\boldsymbol{\theta} = \boldsymbol{M\theta},$$

with $\boldsymbol{\phi}$, $\boldsymbol{\varphi}$, $\boldsymbol{\psi}$ the basis used in the representation of the functional data $X(t)$ and the functional coefficient $\beta(t,T)$ respectively with $\boldsymbol{a} = (a_1, \dots, a_p)'$ and $\boldsymbol{\theta} = (b_{11}, b_{12}, \dots, b_{1r}, b_{21}, \dots, b_{qr})'$ their respective coefficients, and $\boldsymbol{M} = (\boldsymbol{\varphi} \otimes \boldsymbol{\psi})$ where $\otimes$ represents the Kronecker product.

We use B-spline basis for both representations, in case of the functional coefficient a two dimensional base is used resulting of the Kronecker product of two marginal one-dimensional basis. With these representations our model is transform into a multivariate regression model:

$$\begin{aligned}\boldsymbol{\eta} &= \boldsymbol{\alpha} + \boldsymbol{Z\gamma} + \tfrac{1}{\boldsymbol{T}}\textstyle\int_T X(t)\beta(t,T)dt + log\left(\tfrac{\boldsymbol{T}}{365}\right)\\ &= \boldsymbol{Z\gamma} + \boldsymbol{A}\left(\tfrac{1}{\boldsymbol{T}}\textstyle\int_T \boldsymbol{\phi}'\boldsymbol{M}dt\right)\boldsymbol{\theta} + log\left(\tfrac{\boldsymbol{T}}{365}\right)\\ &= \boldsymbol{Z\gamma} + \boldsymbol{A\Psi\theta} + log\left(\tfrac{\boldsymbol{T}}{365}\right) = \boldsymbol{Z\gamma} + \boldsymbol{B\theta} + log\left(\frac{\boldsymbol{T}}{365}\right)\end{aligned}$$

$$\text{with } \boldsymbol{A} = \begin{pmatrix} \hat{\boldsymbol{a}}_1' & 0 & \dots & 0 \\ 0 & \hat{\boldsymbol{a}}_2' & 0 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \hat{\boldsymbol{a}}_N' \end{pmatrix}_{N\times Np} \quad \text{and } \boldsymbol{\Psi} = \begin{pmatrix} \frac{1}{\boldsymbol{T}}\int_T \boldsymbol{\phi}_1'\boldsymbol{M_1} \\ \frac{1}{\boldsymbol{T}}\int_T \boldsymbol{\phi}_2'\boldsymbol{M_2} \\ \vdots \\ \frac{1}{\boldsymbol{T}}\int_T \boldsymbol{\phi}_N'\boldsymbol{M_N} \end{pmatrix}_{Np\times qr}$$

where $\boldsymbol{\phi_i}$ and $\hat{\boldsymbol{a}}_i$ are the basis and coefficients used for the representation of the functional variable $X_i(t)$ and $\boldsymbol{M}_i = (\boldsymbol{\varphi}_i \otimes \boldsymbol{\psi}_i)$, $i = 1, \dots, 110$. $\boldsymbol{\Psi}$ is the matrix of inner products, the calculation of this matrix is one of the novelties of our work. One of the mayor difficulties for its calculation is the fact that the product is between a one-dimensional base and a two-dimensional one. We perform the integration in only one of these dimensions while maintaining the proper bi-dimensional structure. As far as we know this type of inner product has not been done before.

Then for the estimation of the model coefficients we will use Penalized Maximum Likelihood, particularly an anisotropic two dimensional penalization will be used, allowing us to control the smoothness of the functional coefficient independently for each dimension.

$$L_p(\boldsymbol{\theta}, \boldsymbol{y}) = L(\boldsymbol{\theta}, \boldsymbol{y}) - \tfrac{1}{2}\boldsymbol{\theta'P\theta}$$

with $L(\boldsymbol{\theta}, \boldsymbol{y})$ the Poisson likelihood, $\boldsymbol{P} = \lambda_t(\boldsymbol{I}_r \otimes \boldsymbol{D}_2^{'q}\boldsymbol{D}_2^q) + \lambda_T(\boldsymbol{I}_q \otimes \boldsymbol{D}_2^{'r}\boldsymbol{D}_2^r)$ the penalty matrix and $\boldsymbol{D}_2^{'q}$,$\boldsymbol{D}_2^{'r}$ matrices of second order difference of adjacent coefficients of $\boldsymbol{\theta}$.

In order to estimate the smoothing parameters λ_t and λ_T jointly with the rest of the parameters in the model, we reparameterize our model as a mixed model. Therefore, we are in the context of Generalized Linear Mixed Models (GLMMs) and use Penalized Quasi-Likelihood [1] for parameter estimation. To speed up computations we make use of the SOP(Separation of Overlapping Penalties) algorithm [4].

3. Results

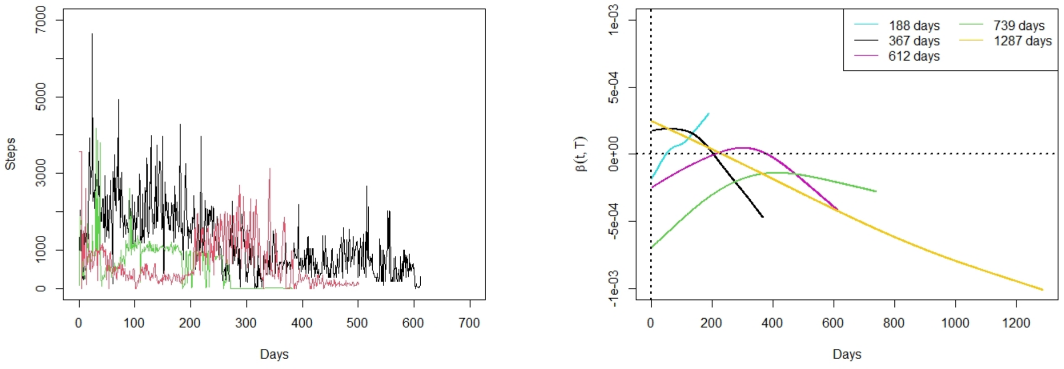


Figure 1: Number of daily steps of 3 patients of the TELEPOC study (left). Curve $\beta(t, T_i)$ for patients with T_i days in the study. (right)

We use the features of the functional parameter $\beta(t, T)$ to understand the relationship between physical activity and the annual ratio of hospitalizations due to COPD. A negative value of $\beta(t, T)$ will imply a positive influence of the physical activity in the reduction of the annual ratio of hospitalizations, on the other hand, a positive value of $\beta(t, T)$ is interpreted as physical activity not helping to reduce the rate of hospitalizations.

Figure 1 (left) shows physical activity patterns of three individuals and Figure 1 (right) shows the trajectories of $\beta(t, T)$ for patients that carried out physical activity for different length periods. The curves present a common feature: doing physical activity regularly during more than 6 months help to reduce the mean number of hospitalizations due to COPD, this conclusion is made based on the fact that all the curves that represent more than 6 months of physical activities end below the horizontal dashed zero line, meaning a favorable influence of the physical activity in the number of hospitalizations due to COPD.

All patients that performed physical activity for more than 8 month eventually turn its influence as favorable in the reduction of the annual ratio of hospitalizations. For patients doing physical activity during 1 year we see how the influence become favorable after, approximately, 200 days. For patients doing physical activity between 1.5 and 2 years we see how, despite an initial unfavorable tendency, after 400 days the influence turns favorable and continues to be so for their remaining time in the study. Finally, for the people performing physical activity for more than 3.5 years, we see how the influence of the physical activity turn favorable over the day 200 and continues to be so for the rest of their time in the study.

Also some non-functional covariates were significant in our model having the following effect: People that had been hospitalize before are more prone to suffer new hospitalizations, also women are more susceptible to be hospitalize than men, depressive symptomatology rise the annual rate of hospitalizations and anxious symptomatology reduces it.

4. Conclusions and future work

We have proposed a new methodology to fit a variable domain functional regression model based on penalized B-splines approximation and their mixed model representation, using the SOP method for the estimation of the functional coefficients and the penalization parameters. This methodology is then applied to the TELEPOC data set where we conclude that regular physical activity helps to reduce the mean number of hospitalizations of COPD patients. We are currently working on the extension of this methodology for the case of more than one functional covariant and the incorporation of adaptive smoothing to our methodology.

5. Acknowledgments

This work is supported by the grant ID2019-104901RB-I00 from the Spanish Ministry of Science, Innovation and Universities MCIN/AEI/10.13039/501100011033.

6. Bibliography

[1] Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear

mixed models. Journal of the American statistical Association, 88(421), 9-25.

[2] Esteban, C., Moraza, J., Iriberri, M., Aguirre, U., Goiria, B., Quintana, J. M., Capelastegui, A. (2016). Outcomes of a telemonitoring-based program (telEPOC) in frequently hospitalized COPD patients. International journal of chronic obstructive pulmonary disease, 11, 2919.

[3] Gellar, Jonathan E. and Colantuoni, Elizabeth and Needham, Dale M. and Crainiceanu, Ciprian M. (2014). *Variable-domain functional regression for modeling ICU data*. Journal of the American Statistical Association, 109, 508 1425-1439. Taylor & Francis

[4] Rodríguez-Álvarez, M. X. and Durban, M. and Lee, D-J. and Eilers, Paul. (2019). *On the estimation of variance parameters in non-standard generalised linear mixed models: Application to penalised smoothing*. Statistics and Computing, 29, 3 483-500. Springer

Spatio-temporal quantile autoregression for detecting changes in daily temperature in northeastern Spain

Jorge Castillo-Mateo[1] Alan E. Gelfand[2] Jesús Asín[1] Ana C. Cebrián[1]

[1]Department of Statistical Methods, University of Zaragoza, Zaragoza, Spain
[2]Department of Statistical Science, Duke University, Durham NC, USA

Abstract

We analyze quantile changes in point-referenced daily temperature in a region of northeastern Spain. We consider regression through asymmetric Laplace (AL) errors in the context of a very flexible mixed effects autoregressive model, introducing two temporal scales and four spatial processes. Moreover, while the autoregressive model yields conditional quantiles, we demonstrate how to extract marginal quantiles with the AL specification.

Keywords: asymmetric Laplace, Bayesian hierarchical model, Markov chain Monte Carlo

1. Introduction

Climate change may lead to changes in several aspects of the distribution of climate variables, including changes in the mean, increased variability, and extreme events. Rather than restricting the analysis to changes in the mean response, quantile regression (QR) permits analysis of several features of the response distribution, including the median but also the extremes. There are two modeling approaches for QR in the literature. The first, usually called *multiple* QR, follows the original ideas by Koenker and Bassett [1], offers a separate regression model for each of the quantiles of interest, and inference proceeds by minimizing a check loss function or assuming asymmetric Laplace (AL) errors. The second approach, usually called *joint* QR, specifies an appropriate joint model for all quantiles [2]. Certainly, QR has received great attention in the statistical literature, including spatial QR models (see, e.g., Lum and Gelfand [3]). Most of these spatial works assume that the data have no temporal dependence. A notable exception is the work of Reich [4], although its model is not autoregressive in time.

Our contribution is to consider QR in the context of a complex spatio-temporal model. This model specifies temporal dependence through autoregression, adopting two time scales, and introduces spatial dependence through Gaussian processes (GPs). Next, we offer an attractive approach to obtain marginal quantiles at daily scale. The analyses consider daily maximum temperature (°C) data at $n = 18$ sites around the Comunidad Autónoma de Aragón provided by the Agencia Estatal de Meteorología (AEMET) in northeastern Spain. The data are available at a daily scale from 1956 to 2015, but we focus the analyses on the warm months from May 1 to September 30. The region presents a wide variety of climate conditions in a relatively small area, bringing interest in studying it and challenge in modeling it. Our interest is in extreme heat so, while our model is applicable to arbitrary quantiles, our primary inferential focus is on the $\tau = 0.95$ quantile.

2. Methodology

2.1. The space-time model

For the analysis of daily maximum temperature, we propose the following spatio-temporal quantile autoregression (QAR) model where each quantile is modeled separately. Letting $Y_{t\ell}(\mathbf{s})$ denote the daily maximum temperature for day ℓ ($\ell = 2, \ldots, L$) of year t ($t = 1, \ldots, T$) at location $\mathbf{s}$ ($\mathbf{s} \in \mathcal{D}$ our study region), $\tau \in (0, 1)$ denote a quantile order and $Q_{Y_{t\ell}(\mathbf{s})}(\tau \mid Y_{t,\ell-1}(\mathbf{s}))$ denote the τ conditional quantile of $Y_{t\ell}(\mathbf{s})$ given $Y_{t,\ell-1}(\mathbf{s})$. Then,

$$\begin{gathered} Y_{t\ell}(\mathbf{s}) = Q_{Y_{t\ell}(\mathbf{s})}(\tau \mid Y_{t,\ell-1}(\mathbf{s})) + \epsilon^{\tau}_{t\ell}(\mathbf{s}) = q^{\tau}_{t\ell}(\mathbf{s}) + \rho^{\tau}(\mathbf{s}) \left(Y_{t,\ell-1}(\mathbf{s}) - q^{\tau}_{t,\ell-1}(\mathbf{s})\right) + \epsilon^{\tau}_{t\ell}(\mathbf{s}), \quad (1) \\ q^{\tau}_{t\ell}(\mathbf{s}) = \beta^{\tau}_0 + \alpha^{\tau} t + \beta^{\tau}_1 \sin(2\pi\ell/365) + \beta^{\tau}_2 \cos(2\pi\ell/365) + \beta^{\tau}_3 \operatorname{elev}(\mathbf{s}) + \gamma^{\tau}_t(\mathbf{s}). \end{gathered}$$

Here, $q^{\tau}_{t\ell}(\mathbf{s})$ contains fixed and random effects. The *fixed effects* are given by β^{τ}_0, a global intercept, $\alpha^{\tau} t$, a global long-term linear trend, sin and cos terms that provide the annual seasonal component, and $\operatorname{elev}(\mathbf{s})$, the elevation at $\mathbf{s}$. The *random effects* given by $\gamma^{\tau}_t(\mathbf{s}) = \beta^{\tau}_0(\mathbf{s}) + \alpha^{\tau}(\mathbf{s})t + \psi^{\tau}_t + \eta^{\tau}_t(\mathbf{s})$ capture space-time dependence through GPs. In particular, $\beta^{\tau}_0(\mathbf{s}) \sim GP(0, C(\cdot; \sigma^{2,\tau}_{\beta_0}, \phi^{\tau}_{\beta_0}))$ and $\alpha^{\tau}(\mathbf{s}) \sim GP(0, C(\cdot; \sigma^{2,\tau}_{\alpha}, \phi^{\tau}_{\alpha}))$ provide local adjustments to the intercept and the long-term linear trend, and $C(\cdot; \sigma^2, \phi)$ is the exponential covariance function. In addition, $\psi^{\tau}_t \sim$ i.i.d. $N(0, \sigma^{2,\tau}_{\psi})$ provides annual intercepts and $\eta^{\tau}_t(\mathbf{s}) \sim$ i.i.d. $N(0, \sigma^{2,\tau}_{\eta})$ provides local annual intercepts. We also specify $\rho^{\tau}(\mathbf{s})$ spatially varying to capture spatial autoregession dependence through $Z^{\tau}_{\rho}(\mathbf{s}) = \log\{(1 + \rho^{\tau}(\mathbf{s}))/(1 - \rho^{\tau}(\mathbf{s}))\} \sim GP(Z^{\tau}_{\rho}, C(\cdot; \sigma^{2,\tau}_{\rho}, \phi^{\tau}_{\rho}))$.

The error term is $\epsilon^{\tau}_{t\ell}(\mathbf{s}) \sim$ ind. $AL(0, \sigma^{\tau}(\mathbf{s}), \tau)$. The AL distribution is characterized by location, scale, and asymmetry parameters, μ, σ, τ; by setting $\mu = 0$ to ensure $P(\epsilon \leq 0) = \tau$, the density of $\epsilon \sim AL(0, \sigma, \tau)$ is written as $f(\epsilon) = \tau(1-\tau)\sigma \exp\{-\sigma\epsilon[\tau - \mathbf{1}(\epsilon < 0)]\}$. A convenient strategy for generating ϵ's is to use the following representation, $\epsilon = \sqrt{\frac{2U}{\sigma^2\tau(1-\tau)}}Z + \frac{1-2\tau}{\sigma\tau(1-\tau)}U$, where $Z \sim N(0, 1)$ and $U \sim Exp(1)$. So, $\epsilon \mid \sigma, U$ is normally distributed enabling us to use all of the familiar Gaussian theory. In the same manner as above, we specify $\sigma^{\tau}(\mathbf{s})$ to capture spatial scale dependence through $Z^{\tau}_{\sigma}(\mathbf{s}) = \log\{\sigma^{\tau}(\mathbf{s})\} \sim GP(Z^{\tau}_{\sigma}, C(\cdot; \sigma^{2,\tau}_{\sigma}, \phi^{\tau}_{\sigma}))$.

Prior distributions Model inference is implemented in a Bayesian framework. Recall that the model adopts a conditional AL distribution for all $Y_{t\ell}(\mathbf{s})$, and this distribution can be expressed as normal when it is conditioned on $U^{\tau}_{t\ell}(\mathbf{s}) \sim Exp(1)$. To complete the model we specify diffuse and, when available, conjugate priors such as normal and inverse gamma for all model parameters.

Model fitting We develop a Metropolis-within-Gibbs algorithm to obtain Markov chain Monte Carlo (MCMC) samples from the joint posterior distribution. In particular, we derive full conditional distributions for each of the parameters, including the $n \times T \times (L-1)$ reparameterized latent exponential variables $\xi^{\tau}_{t\ell}(\mathbf{s}) = U^{\tau}_{t\ell}(\mathbf{s})/\sigma^{\tau}(\mathbf{s})$.

Model assessment through cross-validation We take up model assessment in the context of performance across the L days within year, T years and n locations. That is, we are not implementing model comparison; rather, we are looking at local and global adequacy of the model. In particular, we employ leave-one-out cross-validation and the conditional quantiles are obtained using one-step ahead prediction. We use the following metrics: (i) a version of the $R^1(\tau)$ by Koenker and Machado [5], and (ii) the probability $p(\tau)$ that an observation is less than the conditional quantile. In-sample $R^1(\tau)$ is between 0 and 1, where a value closer to 1 indicates a better fit. The target for $p(\tau)$ is proximity to τ. Analogous versions of these measurements without averaging over days, years, or sites have also been considered.

2.2. Marginal quantiles

We present the general strategy for extracting a marginal quantile from a conditional quantile *after* fitting the conditional quantile model. Considering expression (1), it is attractive to think about $q_{t\ell}^{\tau}(\mathbf{s})$ as a version of a marginal τ quantile for $Y_{t\ell}(\mathbf{s})$. However, $P(Y_{t\ell}(\mathbf{s}) \leq q_{t\ell}^{\tau}(\mathbf{s})) \neq \tau$. So, we seek an additive adjustment to $q_{t\ell}^{\tau}(\mathbf{s})$ that adjusts this probability to τ.

To present the idea in its simplest form, we ignore space and years and suppress the superscript τ in the parameters. We have $Y_\ell = q_\ell + \rho(Y_{\ell-1} - q_{\ell-1}) + \epsilon_\ell$ where the $\epsilon_\ell \sim$ i.i.d. $AL(0, \sigma, \tau)$. In this notation, $Q_{Y_\ell}(\tau \mid Y_{\ell-1}) = q_\ell + \rho(Y_{\ell-1} - q_{\ell-1})$ is the τ quantile of the QAR. For convenience, write this model as $W_\ell = \rho W_{\ell-1} + \epsilon_\ell$ with $W_\ell = Y_\ell - q_\ell$. Upon substitution, we have $W_\ell = \rho^\ell W_0 + \sum_{j=0}^{\ell-1} \rho^j \epsilon_{\ell-j}$. We want the τ quantile of W_ℓ, call it $d_\ell^\tau(\rho, \sigma)$, so that $W_\ell - d_\ell^\tau(\rho, \sigma)$ has 0 as its τ quantile and therefore Y_ℓ has $q_\ell + d_\ell^\tau(\rho, \sigma)$ as the τ marginal quantile. Using the conditional normal form for ϵ_ℓ and defining $\tilde{\epsilon}_\ell \equiv \sum_{j=0}^{\ell-1} \rho^j \epsilon_{\ell-j}$, we have

$$\tilde{\epsilon}_\ell \mid \rho, \sigma, U_\ell, U_{\ell-1}, \ldots, U_1 \sim N\left(\frac{1-2\tau}{\sigma\tau(1-\tau)} \sum_{j=0}^{\ell-1} \rho^j U_{\ell-j}, \frac{2}{\sigma^2\tau(1-\tau)} \sum_{j=0}^{\ell-1} \rho^{2j} U_{\ell-j}\right).$$

Though $\tilde{\epsilon}_\ell$ does not have an AL distribution we can find its τ quantile. For any d, we seek

$$P(\tilde{\epsilon}_\ell < d \mid \rho, \sigma) = \int \int \cdots \int P(\tilde{\epsilon}_\ell < d \mid \rho, \sigma, \{U_j : j = 1, 2, \ldots, \ell\})[\{U_j\}] dU_1 dU_2 \cdots dU_\ell.$$

But given $\{U_j : j = 1, 2, \ldots, \ell\}$, we have the distribution for $\tilde{\epsilon}_\ell$ above. In fact, we can do a Monte Carlo integration to calculate $P(\tilde{\epsilon}_\ell < d \mid \rho, \sigma)$ by generating many sets $\{U_j : j = 1, 2, \ldots, \ell\}$, all i.i.d., all distributed as $Exp(1)$. Then, using a simple search, we can find $d_\ell^\tau(\rho, \sigma)$. In our modeling setting we can create the posterior distribution of the τ marginal quantile for any year, day, and site. In the sequel, we denote this as $\tilde{q}_{Y_{t\ell}(\mathbf{s})}(\tau) \equiv q_{t\ell}^\tau(\mathbf{s}) + d_{t\ell}^\tau(\rho^\tau(\mathbf{s}), \sigma^\tau(\mathbf{s}))$.

Kriging marginal quantiles These marginal quantiles can be kriged over a spatial region for any τ, year, and day within year to reveal the temperature *quantile surface*; i.e., we can obtain the posterior distribution of $\tilde{q}_{Y_{t\ell}(\mathbf{s}_0)}(\tau)$ at any new site $\mathbf{s}_0$. If we obtain this for a sufficiently spatially resolved grid, we can obtain the posterior mean at each point and "see" the posterior τ quantile surface for the given day within year.

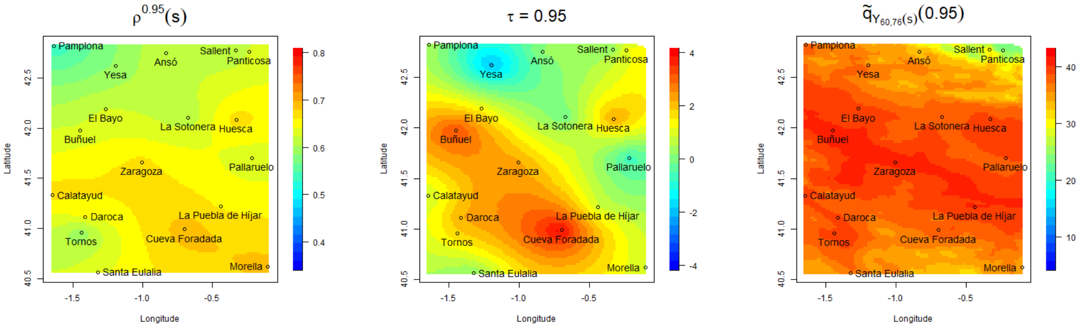


Figure 1: Left: Spatially varying autorregression coefficients. Center: Difference in °C between marginal quantiles of the last and the first decade. Right: Marginal 0.95 quantile on July 15, 2015.

3. Results

Here, we present a brief summary of results for $\tau = 0.95$, although other quantiles have been fitted, obtaining remarkable differences. In summary, model assessment yields $p(0.95) = 0.944$ and $R^1(0.95) = 0.442$. The daily, annual and local measurements also show the good assessment of the model. Results of model fitting yield Figure 1 that shows maps of (left) $E(\rho^{0.95}(\mathbf{s}) \mid data)$, (center) $E\left(\sum_{t\in D6} \tilde{q}_{Y_{t\ell}(\mathbf{s})}(0.95) - \sum_{t'\in D1} \tilde{q}_{Y_{t'\ell}(\mathbf{s})}(0.95) \mid data\right)/10$ where $D1$ is the first decade (1956–1965) and $D6$ the last (2006–2015), and (right) $E(\tilde{q}_{Y_{60,75}(\mathbf{s})}(0.95) \mid data)$. First, $\rho^{0.95}(\mathbf{s})$ shows a strong autoregression and varies spatially from 0.53 to 0.69. Second, warming is general comparing the extremes of both decades, exceeding 3°C in the southwest, but cooling patterns also appear in the northwest. Finally, marginal quantiles enjoy direct interpretation, whereas conditional quantiles require the previous day's temperature. The temperature range for this marginal quantile goes from 23.3°C to 41.1°C.

4. Summary and future work

This article develops a modeling approach to predict a specific quantile in a spatio-temporal framework. We have specified a spatial autoregressive model on a daily scale using the AL distribution for the errors. The model enables spatial autoregression at a daily scale that captures serial correlation and facilitates assessment of persistence. Although the model gives conditional quantiles, we offer an attractive approach to obtain marginal quantiles at daily scale. Posterior inference to evaluate distributional changes between marginal quantiles is straightforward.

Future work will consider addressing quantile crossing using the reparametrization by Yang and Tokdar [2]. This will enable comparison of long-term trends between quantiles jointly. Additionally, our modeling approach could be useful in other environmental analyses such as pollution exposure or biological experimental data to identify distributional changes over space and time.

5. Acknowledgments

This work was partially supported by the Ministerio de Ciencia e Innovación under Grant PID2020-116873GB-I00; Gobierno de Aragón under Research Group E46_20R: Modelos Estocásticos; and JC-M was supported by Gobierno de Aragón under Doctoral Scholarship ORDEN CUS/581/2020.

6. Bibliography

[1] Koenker R., and Bassett G. (1978). Regression quantiles. *Econometrica*, **46**: 33–50.

[2] Yang Y., and Tokdar S. T. (2017). Joint estimation of quantile planes over arbitrary predictor spaces. *Journal of the American Statistical Association*, **112**: 1107–1120.

[3] Lum K., and Gelfand A. E. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis*, **7**: 235–258.

[4] Reich B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**: 535–553.

[5] Koenker R., and Machado J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, **94**: 1296–1310.

Estimation of the area under the ROC curve with complex survey data

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

[1]amaia.iparragirre@ehu.eus, Departamento de Matemáticas, Universidad del País Vasco (UPV/EHU)

[2]irantzu.barrio@ehu.eus, Departamento de Matemáticas, Universidad del País Vasco (UPV/EHU); BCAM - Basque Center for Applied Mathematics

[3]inmaculada.arostegui@ehu.eus, Departamento de Matemáticas, Universidad del País Vasco (UPV/EHU); BCAM - Basque Center for Applied Mathematics

Abstract

In complex survey data, each observation has assigned a sampling weight, which indicates the number of units that it represents in the population. Increasingly, prediction models are being developed from complex survey data. Therefore, the goal of this work is to propose a methodology that considers the sampling weights in the estimation process of the area under the receiver operating characteristic curve.

Keywords: AUC, complex survey data, sampling weights.

1. State of the Art

Complex survey data are gaining popularity in a number of fields, including but not limited to social and health sciences. This type of data are collected from a finite population, concerned to be studied, by some complex sampling design, such as stratification or clustering. One of the differences between complex survey data and simple random samples is that, in the first, each sampled observation has assigned a sampling weight, which indicates the number of units that this observation represents in the finite population. Therefore, the straightforward application of the most commonly applied statistical techniques, which are typically designed to be applied to simple random samples, is usually not suitable for complex survey data.

Among other purposes, complex survey data are increasingly used to develop prediction models. In this work, we focus on logistic regression models to predict dichotomous response variables given a collection of covariates. If these models are fitted in order to be applied for prediction purposes, then it is important for these models to have a good discrimination ability, in order to guarantee their usefulness and accuracy when applied into new individuals. In addition, the estimate of the discrimination ability should be unbiased.

For logistic regression models, the discrimination ability is usually measured by means of the area under the receiver operating characteristic (ROC) curve (AUC). The goal of this study is two-fold. In the first place, we propose a new methodology for estimating the AUC of logistic regression models considering the sampling weights. Secondly, we evaluate the behaviour of the proposed approach by means of a simulation study.

2. Methods

Let $\mathbf{X}$ be the vector of covariates and Y the dichotomous response variable, which takes the value $Y = 1$ for the units with the characteristic of interest (events), and $Y = 0$ otherwise (non-events). Let $p(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$ indicate the conditional probability of event for an individual given the values of its vector of covariates $\mathbf{x}$. Let $\vec{\beta}$ indicate the vector of regression coefficients. Then the specific form of the logistic regression model is:

$$p(\mathbf{x}) = \frac{e^{\vec{\beta}^T \mathbf{x}}}{1 + e^{\vec{\beta}^T \mathbf{x}}}. \tag{1}$$

Each individual can be classified based on its probability of event and a threshold c as event (if $p(\mathbf{x}) \geq c$) or non-event ($p(\mathbf{x}) < c$). However, this classification may or may not be correct depending on the selected threshold. The correct classifications, based on a particular threshold, are usually quantified by sensitivity $(Se(c))$ and specificity $(Sp(c))$ parameters, which are defined as the probabilities of classifying correctly the events and non-events, respectively. The discrimination ability of a logistic regression model is usually evaluated by means of the AUC, where the ROC curve is defined as the set of pairs of $1 - Sp(c)$ and $Se(c)$ across all the possible threshold values c. The AUC ranges from 0.5 (an uninformative model) to 1 (a perfect model in terms of discrimination).

Let $U = \{1, ..., N\}$ be a finite population for which N realizations of the set of random variables $(Y, \mathbf{X})$ are associated, i.e., $\{(y_i, \mathbf{x}_i)\}_{i=1}^{N}$. Let S be a sample of n observations drawn from the finite population U by some complex sampling design. The sampling weights associated to each sampled unit $i \in S$ are denoted as w_i. Let S_0 and S_1 indicate the subsamples of sizes n_0 and n_1 of non-event and event individuals, respectively. In this context, the regression coefficients are usually estimated maximizing the pseudo-likelihood function [1] defined in eq. (2):

$$PL(\vec{\beta}) = \prod_{i \in S} p(\mathbf{x}_i)^{y_i w_i} \left(1 - p(\mathbf{x}_i)\right)^{(1-y_i) w_i}. \tag{2}$$

Let $\widehat{\vec{\beta}}$ indicate the vector of estimated regression coefficients and $\hat{p}_i = \hat{p}(\mathbf{x}_i)$ the corresponding estimated probabilities of event, $\forall i \in S$. The AUC of the fitted model is usually estimated as described in eq. (3), by means of the Mann-Whitney U-statistic [2]:

$$\widehat{AUC} = \frac{1}{n_0 \cdot n_1} \sum_{j \in S_0} \sum_{k \in S_1} \left[I(\hat{p}_j < \hat{p}_k) + 0.5 I(\hat{p}_j = \hat{p}_k) \right], \tag{3}$$

where $I(\cdot)$ denotes the indicator function.

However, we believe that in the context of complex survey data, if we estimate the AUC of the fitted model based on eq. (3), which was designed to be applied in simple random samples and does not consider the sampling weights, then biased estimates could be obtained. For this reason, we propose a new estimator for the AUC which considers the sampling weights. This proposal is described in Section 2.1 below.

2.1. Proposal for estimating the AUC for complex survey data

As mentioned above, there are two ways of thinking about the AUC, and therefore, in this work we have considered two estimators: a) based on the definition of the AUC as the area under the ROC curve; and b) by considering the estimation of the empirical AUC based on the Mann-Whitney U-statistic.

a) Empirical AUC as the area under the ROC curve.

As mentioned above, the AUC is defined as the area under the ROC curve. Therefore, in this first approach, we propose to estimate the ROC curve considering the sampling weights, as follows:

$$\widehat{ROC}_w(\cdot) = \left\{(1 - \widehat{Sp}_w(c),\ \widehat{Se}_w(c)),\ c \in (0,1)\right\}, \tag{4}$$

for which specificity and sensitivity parameters are estimated by means of the sampling weights:

$$\widehat{Sp}_w(c) = \frac{\sum_{j\in S_0} w_j \cdot I(\hat{p}_j < c)}{\sum_{j\in S_0} w_j} \quad ; \quad \widehat{Se}_w(c) = \frac{\sum_{k\in S_1} w_k \cdot I(\hat{p}_k \geq c)}{\sum_{k\in S_1} w_k}. \tag{5}$$

We propose to calculate the area under $\widehat{ROC}_w(\cdot)$ in order to estimate the AUC. Let us denote as $\hat{A}$ the AUC estimated based on this approach.

b) Based on the Mann-Whitney U-statistic.

In this second approach we propose to incorporate the sampling weights into the Mann-Whitney U-Statistics, as follows:

$$\widehat{AUC}_w = \frac{\sum_{j\in S_0}\sum_{k\in S_1} w_j w_k \left[I(\hat{p}_j < \hat{p}_k) + 0.5 \cdot I(\hat{p}_j = \hat{p}_k)\right]}{\sum_{j\in S_0}\sum_{k\in S_1} w_j w_k}. \tag{6}$$

We have demonstrated mathematically that both approaches are equivalent (i.e., $\hat{A} = \widehat{AUC}_w$), and therefore, we will obtain exactly the same estimates in both ways. We have also shown that this proposal is equivalent to one of the definitions given in Yao et al. (2015) [3].

3. Simulation study and results

This simulation study aims, on the one hand, to analyze whether the method we propose provides unbiased results for the estimation of the AUC, and on the other hand, to compare these estimates with the ones we would obtain by ignoring the sampling weights. For this purpose, a finite population was simulated and sampled $R = 500$ times following a complex sampling design. In each of these samples, a model was fitted and its AUC was estimated, considering and/or not the sampling weights. In addition, we are interested in estimating the discrimination ability of each model in the finite population. Therefore we extend the models to the population and estimate the AUCs, which will be defined as the true AUCs. Finally, we compare the unweighted and weighted estimates of the AUC to the true AUCs. This process is described below. For $r = 1, \ldots, R$:

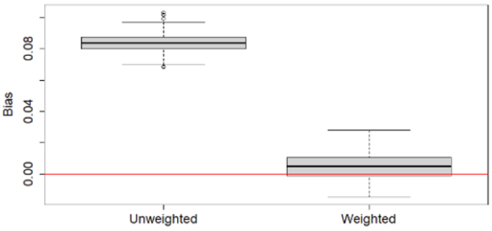


Figure 1: Bias of the weighted and unweighted estimates, $\forall r = 1, \ldots, R$.

	Estimates	Bias
Method	Mean (sd)	Mean (sd)
Unweighted	0.8043 (0.0053)	0.0838 (0.0057)
Weighted	0.7255 (0.0081)	0.0049 (0.0082)
	true AUC	0.7206 (0.0028)

Table 1: Mean and standard deviation (sd) of the estimates and bias for the unweighted and weighted methods across $R = 500$ samples.

Step 1. Obtain a sample $S^r \subset U$ with the corresponding sampling weights $w_i^r, \forall i \in S^r$.

Step 2. Fit the model to S^r and obtain $\widehat{\vec{\beta}^r}$ (eq. (2)) and $\hat{p}_i^r$, $\forall i \in S^r$.

Step 3. Estimate the AUCs by eq. (3) and eq. (6), respectively: $\widehat{AUC}^r_{unw}$ and $\widehat{AUC}^r_{w}$.

Step 4. By means of $\widehat{\vec{\beta}^r}$ estimate the probabilities of event for all the units in the finite population, $\hat{p}_i^r, \forall i = 1, \ldots, N$. Estimate the true AUC in the population following eq. (3): $\widehat{AUC}^r_{true}$.

We define the bias as follows for each method $\forall m \in \{unw,\ w\}$ and $\forall r = 1, \ldots, R$:

$$Bias^r_m = \widehat{AUC}^r_m - \widehat{AUC}^r_{true}. \tag{7}$$

Figure 1 and Table 1 depict the results of the simulation study. It can be observed that the unweighted method gives biased estimates overestimating the true AUC considerably, which would lead to wrong conclusions. In contrast, unbiased estimates are obtained by means of the weighted approach. The standard deviation of the weighted estimates is slightly greater than the unweighted ones.

4. Conclusions and further work

In this work an estimator that considers the sampling weights for the estimation of the empirical AUC in the context of complex survey data is proposed. The results obtained in the simulation study suggest the use of the proposed methodology to estimate the empirical AUC in the context of complex survey data, given that it has shown to give better (unbiased) estimates than the original method (biased estimates) which does not consider the sampling weights. As further work, we believe it would be interesting to define and calculate the variance of this estimator.

5. Acknowledgments

This work was financially supported in part by IT1294-19, PID2020-115882RB-I00, KK-2020/00049. The work of AI was supported by PIF18/213.

6. Bibliography

[1] Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 279 – 292.

[2] Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4), 387-415.

[3] Yao, W., Li, Z., & Graubard, B. I. (2015). Estimation of ROC curve with complex survey data. *Statistics in medicine*, 34(8), 1293-1303.

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

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

[*]brodriguez@barcelonabeta.org, Barcelonaβeta Brain Research Center, Pasqual Maragall Foundation.

[¶]These authors contributed equally to this work.

Abstract

Mendelian randomization (MR) was developed for assessing causality using genetic variants in epidemiological research. The objective of this study was to compare several robust MR methods to discern causal effects in epidemiological studies. As a proof of concept, we tested the potential causal role of telomere length (TL) (a well-known biomarker of aging) on life expectancy.

Keywords: Causal inference, epidemiology, genetic association studies

1. Introduction

Mendelian Randomization (MR) approaches were developed to assess causality in observational studies using the fact that genetic variants segregate following Mendel's laws. Estimating how modifiable risk factors or exposures of interest (*i.e.*, X) influence specific diseases or outcomes (*i.e.,* Y) can be challenging. Observational studies suffer from confounding and reverse causation, which impede to discern whether the association between the exposure and outcome is causal or not. MR approaches use genetic variants (Z) known to affect X as surrogates for exposure assessment to overcome problems in traditional epidemiology. Thus, genetic variants serve as instrumental variables (IVs) to obtain unbiased estimators of the exposure-outcome causal relationship [1].

Including multiple genetic variants as IV can increase the statistical power and therefore the precision of MR estimates. Genome-wide association studies (GWAS) of disease biomarkers are providing a prolific source of IV for MR studies, Unfortunately, the limited knowledge of their functional role together with the ubiquitous presence of pleiotropy (i.e., their association with multiple phenotypes) could lead to violation of MR assumptions [2]. Nevertheless, various robust MR methods with different assumptions can be applied to explore the consistency and robustness of MR results [3].

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

2. Methods

The inverse-variance weighted (IVW) method is the most widely used and efficient method in MR analyses [4]. However, it assumes all genetic variants are valid instrumental variables, so it is biased in the presence of unbalanced pleiotropy [3]. Several methods are commonly used to identify pleiotropy and confirm the validity of the IVW estimates, including Cochran’s Q test and MR-Egger intercept test.

Further, robust MR methods are useful either to confirm the consistency of MR results or explore causal associations in the presence of pleiotropy. Despite their lower statistical efficiency, robust estimation methods allow for the presence of pleiotropy and the inclusion of invalid IVs in MR analyses. According to the different assumptions required for consistent estimation, we can classify robust methods in consensus methods (*e.g.*, weighted median and weighted mode), outlier-robust (*e.g.*, MR-PRESSO) and distribution modelling from invalid IVs (*e.g.*, MR-Egger regression) [5].

Mendelian Randomization modelling assumptions

For a genetic variant to be considered as valid IV in MR analysis, it must satisfy the following assumptions: it is associated with the exposure (assumption 1); the genetic variant is not associated with the outcome due to confounding factors (assumption 2) and it has no direct effect on the outcome (assumption 3).

Inverse-variance weighted method

The inverse-variance weighted (IVW) method assumes that the genetic association with the exposure ($\hat{\beta}_{Xj}$) and the genetic association with the outcome ($\hat{\beta}_{Yj}$) are related via a simple linear regression model. Therefore, IVW is equivalent to fitting a weighted regression model where the regression coefficient is an estimate of the causal effect ($\hat{\beta}_{IVW}$) if all genetic variants are valid IVs. In addition, there is no intercept term (α_j) because of the third MR assumption (*i.e.*, no pleiotropy):

$$\hat{\beta}_{Yj} = \hat{\beta}_{IVW}\hat{\beta}_{Xj}Z + \varepsilon\mathrm{j}$$

When using summary statistics from GWAS, IVW combines the Wald ratio causal estimates as the average of all the $\hat{\beta}_{Yj}/\hat{\beta}_{Xj}$ effects weighted by the inverse of their variance:

$$\hat{\beta}_{IVW} = \frac{\sum_j \hat{\beta}_{Yj}\hat{\beta}_{Xj} SE\left(\hat{\beta}_{Y_j}\right)^{-2}}{\sum_j \hat{\beta}_{Xj}^2 SE\left(\hat{\beta}_{Yj}\right)^{-2}} \qquad SE\left(\hat{\beta}_{IVW}\right) = \sqrt{\frac{1}{\sum_j \hat{\beta}_{Xj}^2 SE\left(\hat{\beta}_{Yj}\right)^{-2}}}$$

Cochran's Q statistic, derived from the IVW estimate, can be used to detect heterogeneity due to pleiotropy and invalid IVs.

MR-Egger regression

MR-Egger relaxes the assumption that the average pleiotropic effect is zero by introducing an intercept term in the regression model. MR-Egger estimates rely on the InSIDE assumption (*i.e.*, Instrument Strength Independent of Direct Effect). The assumption states that the pleiotropic effects, α_j, are independently distributed from the genetic associations with the exposure $\hat{\beta}_{Xj}$ [6]. Further, it is possible to detect the presence of directional pleiotropy by testing whether the intercept term differs from zero (the MR-Egger intercept test).

Mendelian Randomization Consensus methods

MR-consensus methods estimate the causal effect based on a summary measure of the distribution of the Wald ratio estimates [5]. These methods provide unbiased estimators even in the presence of invalid IVs. For instance, the weighted median method [7] assumes that the majority (*i.e.*, >50%) of the genetic variants are valid IVs and so it will be less influenced by outlying variants than the IVW estimate. Further, in large samples, causal estimation could be unbiased even when only a plurality of genetic variants are valid IVs. The weighted mode method relies on this plurality valid assumption (*i.e.*, Zero Modal Pleiotropy Assumption) by estimating the causal effect as the most common variant-specific ratio estimate [8].

Outlier detection

The MR-Pleiotropy RESidual Sum and Outlier (MR-PRESSO) global test evaluates overall pleiotropy by comparing the residual sum of squares (RSS) with the expected distance under the null hypothesis of no pleiotropy [9]. Further, it detects outlier IVs by omitting each genetic variant from the analysis in turn and checking whether the RSS decreases [5].

3. Proof of concept

Genetic Instrumental Variables of Telomere Length and summary statistics of Life Expectancy

Summary statistics were obtained from a genome-wide meta-analysis of leukocyte telomere length (TL) including up to 78,592 individuals [10]. We selected summary data for 21 genetic variants associated with TL ($P < 5\times10{-}8$) and the most recent large-scale GWAS on life expectancy (N=75,244) [11].

Mendelian Randomization performance

Genetically predicted longer TL was significantly associated with increased life expectancy according to IVW method with no evidence of heterogeneity (Cochran Q p-value > 0.05) nor directional pleiotropy (intercept test p-value > 0.05) [Table 1].

Pleiotropy-robust methods (*i.e.*, weighted median and weighted mode) also produced similar patterns of effects, further supporting the robustness of the analyses. Interestingly, MR-Egger regression method did not support this association (MR-Egger p-value = 0.113). MR-Egger estimates could be biased in the presence of outliers. However, there was no evidence of heterogeneity due to SNP outliers according to MR-PRESSO global test. MR-Egger is also sensitive to violations of InSIDE assumption; however, it usually leads to increased bias and Type 1 error rate inflation [12]. Nevertheless, MR-Egger may be less efficient if this assumption does not hold [5].

As our main conclusion, we recommend performing robust methods with different assumptions to test the consistency and robustness of MR results in genetic epidemiology research.

Causal inference methods	***β*** **(SE)**	**p-value**
Inverse-variance weighted	0.011 (0.004)	0.010
MR-Egger regression	0.020 (0.013)	0.113
Weighted median	0.013 (0.006)	0.038
Weighted mode	0.021 (0.009)	0.021
Sensitivity methods	***p*****-value**	
Cochran Q, heterogeneity test	0.978	
MR-PRESSO, global test	0.982	
MR-Egger, intercept test	0.462	

Table 1. Mendelian randomization results for the relationship between genetically predicted longer telomere length and life expectancy.

4. Bibliography

[1] Katan, M. (1986) "Apolopoprotein e isoforms, serum cholesterol, and cancer," *The Lancet*, 327(8479), pp. 507–508.

[6] Glymour, M. M., et al. (2012) "Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions," *American Journal of Epidemiology*, 175(4), pp. 332–339.

[3] Burgess, S. et al. (2020) "Guidelines for performing Mendelian randomization investigations," *Wellcome Open Research*, 4, p. 186.

[4] Burgess, S., Butterworth, A. and Thompson, S. G. (2013) "Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data," *Genetic Epidemiology*, 37(7), pp. 658–665.

[5] Slob, E. A. W. and Burgess, S. (2020) "A comparison of robust Mendelian randomization methods using summary data," *Genetic Epidemiology*, 44(4), pp. 313–329.

[6] Burgess, S. and Thompson, S. G. (2017) "Interpreting findings from Mendelian randomization using the MR-Egger method," *European Journal of Epidemiology*, 32(5), pp. 377–389.

[7] Bowden, J. et al. (2016) "Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator," Genetic Epidemiology, 40(4), pp. 304–314.

[8] Hartwig, F. P., et al. (2017) "Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption," *International Journal of Epidemiology*, 46(6), pp. 1985–1998.

[9] Verbanck, M. et al. (2018) "Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases," Nature Genetics, 50(5), pp. 693–698.

[10] Li, C. et al. (2020) "Genome-wide Association Analysis in Humans Links Nucleotide Metabolism to Leukocyte Telomere Length," The American Journal of Human Genetics, 106(3), pp. 389–404.

[11] Pilling, L. C. et al. (2016) "Human longevity is influenced by many genetic variants: evidence from 75,000 UK Biobank participants," 8(3), pp. 547–560.

[12] Burgess, S. and Thompson, S. G. (2017) "Interpreting findings from Mendelian randomization using the MR-Egger method," European Journal of Epidemiology, 32(5), pp.377–389. doi: 10.1007/s10654-017-0255-x.

Parallel session: Survival analysis

Chairperson:

Klaus Langohr, Universitat Politecnica de Catalunya (Spain)

A solution to implement reference-based imputation with time-to-event endpoints through the illness-death multi-state model

Alberto García-Hernandez[1], *Teresa Pérez*[2], *María del Carmen Pardo*[3],
Dimitris Rizopoulos[4]
[1]albega28@ucm.es, Facultad de Estudios Estadísticos, UCM, Madrid, Spain
[2]teperez@estad.ucm.es, Facultad de Estudios Estadísticos, UCM, Madrid, Spain
[3]mcapardo@ucm.es, Facultad de Ciencias Matemáticas, UCM, Madrid, Spain
[4]d.rizopoulos@erasmusmc.nl, Erasmus Medical Center, Rotterdam, The Netherlands

In randomized clinical trials, it is difficult to fully adhere to the intent-to-treat principle if subjects are not followed up after discontinuation of the study treatment. This situation is regarded a censoring-not-at-random (CNAR) problem because the observed data (emerging from the on-treatment period) are regarded insufficient to infer the risk of event after stopping the study medication.

This problem can be addressed using reference-based imputation (RBI) methods. The arguably most popular RBI assumption is the so-called jump to reference (J2R) rule. Using J2R, after treatment discontinuations, subjects in the experimental arm are assumed to lose the benefit observed during the treatment period and switch back to the risk function observed in the reference group.

For time to event endpoints, RBI has been exclusively studied using Rubin's multiple imputation (MI) methodology. Here, we present a fully analytical maximum-likelihood solution. We present this problem as a particular case of the illness-death multi-state model. The transition from the initial state (I) to drug discontinuation (DD) and from the initial state (I) to the event of interest (E) (while on-treatment) is well characterized from the dataset. For the unobserved transition from DD to E, we use well-established RBI assumptions such as J2R. Finally, we calculate the probabilities of interest (i.e., transitioning from I to E regardless of having passed or not by DD) using numerical integration.

Our analytical solution is more efficient than MI and, as opposed to MI, avoids standard error over-estimation.

Keywords: reference-based imputation, jump-to-reference, survival analysis, multi-state modelling.

Non-Markov multistate models applied to a cohort of COVID-19 patients

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

[1]mbesalu@ub.edu, Dept. Genètica, Microbiologia i Estadística, Universitat de Barcelona
[2]lupe.gomez@upc.edu, Dept. d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya-BarcelonaTech (UPC) and Institute of Mathematics of UPC-BarcelonaTech (IMTech)

Multistate models are a convenient way to describe the course of a disease. In multistate Markov models the future evolution only depends on the current state and, although Markov assumption is rarely met in practice, they are often used for simplicity. We explore options of enriching pathway information by means of non-Markov approaches and second order Markov processes.

Second order Markov models assume that the progression of the patients not only depends on the current but also on the preceding state. Estimation and inference entails a $M \times M \times M$ matrix, where M is the number of states. In this work, we define an extended transition probability matrix as M different matrices of order $M \times M$ and we propose nonparametric methods to deal with.

Non-Markov models, not assuming any restriction, allow a comprehensive way of dealing with the different path trajectories. We follow the work of Putter and Spitoni (2018) to estimate the transition probabilities by means of the landmark Aalen-Johansen estimator constructed from a subset of the data consisting of all subjects observed to be in a given state at a given time.

The above methods are explored to analyze a cohort of about 2000 COVID-19 patients from eight hospitals in Catalunya who were hospitalized during the first wave of the pandemic. We have built a multistate model based on 14 transitions among the following seven states after a patient's admission: No severe pneumonia, severe pneumonia, non-invasive mechanical ventilation, invasive mechanical ventilation, severe pneumonia recovery, hospital discharge, and death. Of special relevance is the transition probability to severe pneumonia recovery of patients requiring invasive mechanical ventilation. A second order Markov assumption allows us to estimate this transition probability accounting for whether or not the patient was on previous non-invasive mechanical ventilation. If the interest would be on this transition probability after a week accounting for the whole past history, the landmark Aalen-Johansen estimator could be applied based only on all those patients with invasive mechanical ventilation at day 7.

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

A prediction model for survival in patients with advanced disease: dealing with non-proportional hazards factors.

Peñafiel J [1], Turrillas P[2], Gomez-Batiste X[2], Tebe C [1]

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

[2] pturrillas@gmail.com, xavier.gomez@uvic.cat, University of Vic-Central University of Catalonia (UVIC-UCC), Vic, Barcelona, Spain.

Objective: To develop a clinical prediction model for survival prediction at 24 months in patients with advanced diseases.

Methods: A retrospective cohort of 736 patients was randomly split in a derivation (80%) and validation cohort (20%). 6 survival-related variables were pre-selected by a group of clinical experts (Age / Palliative care needs /Nutritional decline/ Resources use / Main disease / Multimorbility / Functional decline) to be included in a Cox proportional hazard model. The prognostic value of the estimated model was assessed using: AIC, C-statistic of Somers 'Dxy' rank correlation, integrated Brier score for censored observations and the ratio between the observed and the predicted risk with 95% confidence interval. The satisfaction of proportional hazard assumptions was assessed graphically and analytically. Whether the assumptions were not satisfied two strategies were defined: stratified Cox model and a time-dependent Cox model. All analysis were performed using R.

Results: 630 patients were included in the derivation cohort (59.7% women, 80.7-yo, 34% advanced fragility, 33% organ disease, 21% dementia, and 12% cancer). A full adjusted Cox model was estimated. Palliative care needs and main disease presented a severe violation of the proportional hazard assumption in that model. Then a full adjusted Cox model was estimated using palliative care needs as a time-dependent factor. As an alternative, a full adjusted stratified Cox model was estimated using palliative care needs and main disease as stratum a full adjusted stratified Cox model was estimated using palliative care needs and main disease as stratum. Finally, stratified model was simplified reclassifying main disease as cancer vs non-cancer. Performance in estimated models was: AIC 3247.77, 2251.74 and 2733.87 and , AUC[IC95%] 0.718[0.68;0.756], 0.608[0.574;0.642] and 0.612[0.58;0.645] , Brier score 0.134, 0.179 and 0.173 and E/O[IC95%] 0.73 [0.66;0.79], 0.89 [0.78;1.01] and 0.86 [0.75;0.98]). Model 3 was selected due to parsimonious and achievement. The performance in the derivation and validation cohort was similar.

Conclusion: Survival factors in patients with advanced disease seem to not fulfill proportional hazard assumptions. Baseline hazard stratification by those factors seems an easy and helpful solution.

Keywords: Prediction Models, Cox's proportional risks, Palliative care

A semi-Markov multistate model for in-hospital survival to examine the first COVID-19 wave in the Barcelona metropolitan area

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

[1]Department of Statistics and Operations research, Polytechnic University of Catalonia
xavier.piulachs@upc.edu, klaus.langohr@upc.edu, lupe.gomez@upc.edu

[2]Biostatistics Unit, Bellvitge Biomedical Research Institute, Barcelona, Spain
gmolist@idibell.cat, npallares@idibell.cat, ctebe@idibell.cat

[3]Department of Infectious Diseases, Bellvitge University Hospital, Barcelona, Spain.
cgudiol@bellvitgehospital.cat

Focusing on the Barcelona metropolitan area (Spain), the present work aims to re-examine the multi-state process described by the first wave of COVID-19. A dataset with demographic and clinical information on about 2000 hospitalized patients is used, this providing a detailed account of patient disease progression as transitions through a prefixed set of possible clinical states. Specifically, the modeling process considers five transient states (no severe pneumonia, severe pneumonia, non-invasive mechanical ventilation, invasive mechanical ventilation, and recovery from severe pneumonia), and two absorbing states (hospital discharge and death). Moreover, a total of 14 possible transitions between subsequent states are considered. In a multi-state survival setting, a single time scale is usually adopted, for which transition intensities are computed with respect to the overall origin. In some situations, however, it becomes crucial to register whether a patient has gone through a particular intermediate state, as well as the length of stay in a given state. These are two practical scenarios in which patients under study incorporate non-baseline information, so the Markovian hypothesis needs to be relaxed for a proper description of disease after hospitalization. In doing so, a semi-Markov multi-state model is here proposed, built with semi-parametric Cox models that not only account for the baseline effect of transition-specific covariates, but also for the potential impact of the sojourn time at a previous state. Hence, the specific purpose of this study is twofold: 1) to estimate the probability of reaching different stages of disease progression, conditioned on a set of subject features; 2) to assess the effect of the most important prognostic factors on disease progression by combining different time scales. The promising results of this work advocate for the usefulness of the procedure to guide decision-making and allocation of hospital resources for Covid-19 patients.

Keywords: COVID-19 hospitalizations, survival analysis, multi-state model.

1

A multi-state model to analyze hospitalized Covid-19 patients during the first three waves in the Barcelona metropolitan area.

Molist G[1], Piulachs X[2], Alonso G[3], Rombauts A[3], Pallarès N[1], Langohr K[2]

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

[2] xavier.piulachs@upc.edu klaus.langohr@upc.edu GRBIO: Research Group in Biostatistics and Bioinformatics, Universitat Politècnica de Catalunya-BarcelonaTech

[3]gabi.abelenda.alonso@gmail.com, alexander.rombauts@gmail.com Infectious Diseases Department, Bellvitge University Hospital, IDIBELL, University of Barcelona, Spain.

Objective: To study COVID-19 patients' progress and to estimate the probability of attaining different stages of disease progression within a given period, conditionally on a set of subject features in several COVID-19 waves.

Methods: We analyzed data from hospitalized patients within the south metropolitan area of Barcelona throughout the first three pandemic waves. Adults with COVID-19 and without therapeutic limit at admission were included in the analyses. Hospital admission was the model's common initial state for all patients, and the time spanning from then until the occurrence of a particular endpoint was monitored. The possible states were: 1) severe pneumonia; 2) non-severe pneumonia; 3) invasive mechanical ventilation; 4) noninvasive mechanical ventilation; 5) death and 6) hospital discharge. The total number of possible transitions were 14, but a single patient could not experience more than 4 transitions. To describe a stochastic process allowing subjects to move through a set of transitions in continuous irregular times, we used a semi-Markov multi-state model.

Results: We included about 3000 patients with a mean age of 59 years (from 20 to 96) and 42% were women. The most frequent transition was from hospital admission to discharge (60%), followed by a transition from admission to discharge but first passing through the severe pneumonia state (12%). The 30-day mortality was 10.7% (95% CI: 9.59%–11.97%). The multi-state model allowed for accurate predictions of state transitions, and, in addition, for assessment of various clinical profiles at different time points: the estimated 30-day mortality probability for a woman aged < 65 years, oxygen saturation > 95% and no obesity was 19.64% (95% CI: 15.38%–25.07%), whereas the estimation of the same probability for a woman with severe pneumonia, aged ≥ 65 years, with obesity, and oxygen saturation < 95%, was 34.64% (95% CI: 29.63%–40.51%). In men with the same profiles, the estimated 30-day mortality rates were 18.44% (95% CI: 14.13%–24.05%) and 33.31% (95% CI: 28.46%–38.98%), respectively.

Conclusions: Using a multistate approach, we were able to assess relevant clinical profiles in terms of risk and perform an outcome prediction based on the baseline subject status. This methodology can in turn be applied to other epidemics.

Keywords: Multistate models, COVID-19, semi-markov.

Parallel session: Bio-Bayes modelling

Chairperson:

M. Eugenia Castellanos, Universidad Rey Juan Carlos (Spain)

A Bayesian approach to population estimation in capture-recapture models

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 can be used to estimate the population size or the species richness. Sometimes, when trying to estimate the species richness, the only data available refer to species that have been observed only once (n_1) and species that have been observed on multiple occasions (n_{2+}), called incidence data. This is the case of the study of the number of species in the coral reef of Tunku Abdul Rahman Marine Park[1].

The main objective of this research is to estimate species richness and for this it is necessary to estimate the number of undetected species (n_0) or more specifically, to estimate the proportion of undetected species (p_0). The likelihood function for this problem has a simple form, but that parameters are not identifiable[2] and, consequently, a direct frequentist approach is not possible. However, Bayesian methods are suitable to cope with this problem.

Because we are working with proportions and they sum one, we need to use a multivariate prior distribution defined on the simplex. Our first choice, the Dirichlet distributions (including the uniform), are not suitable because they act as ”killer priors” leading to marginal posterior distributions for $\hat{p_0}$ not depending on the observed data. We explore other more suitable prior distributions for obtaining meaningful posterior marginal distributions for $\hat{p_0}$. Specifically, we consider the Scale Dirichlet, the maximum entropy, and uniform and beta conditional distributions.

Beyond this, we consider different prior scenarios for our data by comparing the obtained results. Other different examples of application are also analysed and discussed.

Keywords: population estimation, identifiability, improper priors

References

[1] 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.

[2] Gelfand, A. E., & Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445), 247-253.

Urban greening and gentrification

David Conesa[1,*], *Isabelle Angelowski*[2], *Miguel Ángel Beltrán*[3], *James J.T. Connolly*[2], *Jesua López-Máñez*[4], *Joaquín Martínez-Minaya*[5], *Blanca Sarzo*[1,6]

[1] Universitat de València
[*] david.v.conesa@uv.es
[2] Universitat Autónoma de Barcelona
[3] Fundación Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana
[4] Caixabank
[5] Universitat Politècnica de València
[6] The University of Edinburgh

Although urban greening is universally recognized as an essential part of sustainable and climate-responsive cities, a growing literature on green gentrification argues that new green infrastructure, and greenspace in particular, can contribute to gentrification, thus creating social and racial inequalities in access to the benefits of greenspace and environmental and climate injustice. To date, there is limited quantitative evidence documenting the temporal relationship between new greenspaces and gentrification across entire cities, let alone across various international contexts. In this paper, we employ a Hierarchical spatial Bayesian model to test the green gentrification hypothesis across 28 cities in 9 countries in North America and Europe. Here we show a strong positive and relevant relationship between greening in the 1990s-2000s and gentrification that occurred between 2000-2016 in 17 of the 28 cities. Our results also further refine the conceptualization of gentrification and green gentrification by identifying cities which we call "Lead Green Gentrification" where greening plays a central role in explaining gentrification unlike in other cities where greening plays what we see as an "integrated" or "subsidiary" role. Results also point to the need to put equity at the center of green climate strategies.

Keywords: Bayesian modeling; INLA; Random effects

Reducing the use of fungicides in agriculture by 50% with decision support systems. A Bayesian meta-analysis

Elena Lázaro[1, *], *David Makowski* [2] *and Antonio Vicent*[1]

[1] Centre de Protecció Vegetal i Biotecnologia, Institut Valencià d'Investigacions Agràries (IVIA), 46113 Moncada, Valencia, Spain.

[2] INRAE, Applied Mathematics and Computing Unit (UMR 518) INRAE AgroParisTech Université Paris-Saclay, 75231 Paris, France.

[*]lazaro_ele@gva.es

The Farm to Fork Strategy of the European Green Deal aims to halve the use of pesticides and specially fungicides by 2030. This goal represents a major challenge for the European agriculture, which will be forced to decrease the number of treatment applications in a very short time while maintaining the profitability of cropping systems. In this context, decision support systems (DSSs) have been put forward as tools to substantially reduce fungicide application frequency. In contrast to calendar-based fungicide programs, DSSs allow farmers to schedule fungicide applications based on an observed or predicted risk of disease and thus spray only when necessary.

To assess whether the “50% reduction” is an achievable goal a comparative study to assess disease incidences between DSSs and calendar strategies to the number of spray applications was addressed by using multilevel Bayesian meta-analysis under a beta-binomial mixed-effect regression modelling framework. Model parameters were approximated by means of Hamiltonian Monte Carlo (HMC) simulation methods using the programming *Stan* via its *R* interface, *RStan* and the package *brms*.

Statistical analyses were rigorously conducted to ensure the robustness of the results by exploiting the potential of Bayesian methods and the functionalities of the aforementioned packages: i) several variants of the generic model were fitted in order to investigate the possible effects of some relevant moderator variables; ii) model evaluation and model selection between all meta-analytic models fitted were carried out considering posterior predictive checks and k-fold cross validation procedures, respectively; and iii) a sensitivity analysis in relation to the setting of prior specification was also addressed comparing parameter estimates with frequentist estimates.

Our global meta-analysis which incorporated 80 experiments - conducted worldwide across different geographic areas, crops, fungal pathogens, and fungicide types - shows that DSSs can reduce fungicide treatments by at least 50% without compromising disease control. Specifically, when the number of sprays was halved, the resulting increase in disease incidence was lower when adopting a DSS-based rather than a calendar-based strategy and never exceeded a 5% increase in disease incidence with DSSs. Thus, our results support that DSSs are essential tools for reducing fungicide use while limiting the risks to plant health.

Keywords: *Stan*, systematic-review, Farm to Fork Strategy

Metamodelling of multimodelling: a Bayesian meta predictive-model for Climate Change

Joaquín Martínez Minaya[1], Dae-Jin Lee[2] and Anabel Forte[3]

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

[2]dlee@bcamath.org, Applied Statistics research line, Basque Center for Applied Mathematics (BCAM)

[3]anabel.forte@uv.es, Department of Statistics and Operations Research, University of Valencia (UV)

In many scientific fields and real-life applications, there are many situations where a statistical modelling process can be approached from different perspectives. The problem arises when we have to choose a single approach to perform inference and make predictions. There is a vast literature related to model selection, from the use of measures such as AIC, DIC, or WAIC, till using the Bayes factor to find which model would be the most likely, or using the Bayesian Model Averaging (BMA) which allows us to average them all, allowing to construct a model combination.

In most cases, the prerequisite for applying these techniques is that they all start from the same likelihood. However, we will not always have a model of the same process from the same likelihood, as these models could be from a deterministic nature, for example, in the context of climate change predictions (e.g: via Generalized Circulation Models, GCM). Then, we propose a method to find the best model combination which reflects the nature of a process, using the predictions obtained by the different models, such as the model's complexity does not need to be taken into account.

In this work, the different models used in the prediction would act as covariates attempting to explain the process of interest. Hence, different multimodels, i.e. combinations of such models could be generated, starting from a linear regression. For each generated multimodel, we would obtain a posterior probability. Thus, we could choose the multimodel with the highest probability, or, we could use the BMA approach to combine the multimodels with each other obtaining the metamodel we are looking for. The benefit of our proposal is twofold: first, the combination of the models is done in a simple way without having to resort to the full inclusion of the complex structure of the models. Second, the use of the BMA methodology allows us to have a weight for each of the multimodels and not only for each of them.

We have applied this method to construct a metamodel for predicting rainfall erosivity (also known as R-factor) in the Basque Country under different Climate Change scenarios. The metamodel generated with our approach showed promising results compared with the GCM models set separately based on the predictive performance.

Keywords: Bayes factor, Bayesian Model Averaging, Climate Change, Metamodel

Multivariate analysis of the determinants of quality composts from green waste streams from different origins using Bayesian networks

Xavier Barber[1], *Francisco Guilabert*[2], *María Dolores Pérez-Murcia*[2], *Enrique Agulló*[2],
Francisco Javier Andreu-Rodríguez[2], *Raul Moral*[2], *Maria Ángeles Bustamante*[2]

[1]Centro de Investigación Operativa (CIO-UMH), Universidad Miguel Hernández de Elche
[2]Centro de Investigación e Innovación Agroalimentaria y Agroambiental (CIAGRO-UMH), Universidad Miguel Hernández de Elche

Organic wastes of vegetal origin and, in particular, those coming from sources related to the tourist activity, such as those generated in golf courses and tourist coasts, are a growing concern due to the increase of their production and their inadequate management. The aim is to evaluate the use of different composting strategies for the management of these specific green wastes, such as grass clippings and pruning waste from a golf course and marine plant debris, mainly Posidonia oceanica (Posidonia oceanica L.). For this purpose, two composting scenarios were established: in the first one only green wastes were considered in the composition of the composting mixtures, and in the second one sewage sludge was used as a co-composting agent. The temperature of the piles was controlled and the physicochemical and chemical parameters were also studied throughout the process. All will be analyzed using additive models and Bayesian networks.

Keywords: Compost quality, golf course, Bayesian networks

AMS: 62P12, 62L10