SORT

Statistics and Operations Research Transactions

Volume
49

Number 2, July-December 2025

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Statistics and Operations Research Transactions

Volume 49, Number 2, July-December 2025

eISSN: 2013-8830

www.idescat.cat/sort/

Secretary

Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel. +34 - 93 557.30.76 - 93 557.30.00
E-mail: sort@idescat.cat

Publisher: Institut d'Estadística de Catalunya (Idescat)

eISSN: 2013-8830
DL B-46.085-1977
Key title: SORT
Numbering: 1 (december 1977)
www.idescat.cat/sort/

eISSN: 2013-8830
SORT 49 (2) July-December (2025)

SORT

Statistics and Operations Research Transactions

Coediting institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Universitat Pompeu Fabra
Universitat de Lleida
Universitat Rovira i Virgili
Institut d'Estadística de Catalunya

Supporting institutions

Spanish Region of the International Biometric Society
Societat Catalana d'Estadística
Departament de Recerca i Universitats

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Volume 49
Number 2
July-December 2025
eISSN: 2013-8830

Articles

SORT 49 (2) July-December 2025, 149-178 DOI: 10.57645/20.8080.02.26

A stochastic partial differential equation for Bayesian spatio-temporal modeling of crime

Julia Calatayud[1], Marc Jornet[2], Javier Platero[3] and Jorge Mateu[4]

Abstract

We propose a stochastic partial differential equation to model geo-referenced data in the plane, with spatially correlated noise and a temporal log-normal evolution. Discretization in space permits us to develop the model in a finite-dimensional framework, reducing it to a set of stochastic differential equations coupled by correlated Wiener processes. The correlations are considered time-varying and stochastic, with a transformed log-normal distribution. The final model is framed within a hierarchical structure, and parameter inference is conducted jointly using Bayesian methods. The statistical methodology is illustrated by analyzing crime activity in the city of Valencia, Spain.

MSC: *35R60, 60H10, 62F15, 62M30, 62P25.*

Keywords: *Bayesian inference, crime time series, lattice data, space-time correlation, space-time intensity, stochastic log-Gaussian model, stochastic partial differential equation.*

1. Introduction

The use of differential equations in mathematical criminology started with the study of dynamic properties of urban crime hotspots, with the seminal paper of Short et al. (2008). The degree of attractiveness of each site influences the movement of burglars. The proposed model consists of two coupled reaction-diffusion partial differential equations describing the spatio-temporal evolution of density and attractiveness, giving rise to crime pattern formation. Building on this paper, extensions and further investigations have then been conducted on the dynamics of the reaction-diffusion system; see, for instance, the contributions by Short, Bertozzi and Brantingham (2010a; 2010b); Lloyd

[1] Department of Mathematics, University Jaume I, 12071 Castellón, Spain. calatayj@uji.es

[2] Escuela Superior de Ingeniería y Tecnología, Universidad Internacional de La Rioja, Av. La Paz 137, 26006 Logroño, Spain. marc.jornet@unir.net

[3] Department of Mathematics, University Jaume I, 12071 Castellón, Spain. platero@uji.es

[4] Department of Mathematics, University Jaume I, 12071 Castellón, Spain. mateu@uji.es

Received: May 2024
Accepted: April 2025

and O'Farrell (2013); Kolokolnikov, Ward and and Wei (2014); Tse and Ward (2016); Wang, Wang and Feng (2020); Rodriguez and Winkler (2022) and references therein.

Spatially homogeneous models of crime, based on temporal compartmental systems of ordinary differential equations, seem to begin with McMillon, Simon and and Morenoff (2014). The authors took a population-based approach, rooted in models that have been developed mainly to model the spread of diseases (Brauer, 2008; Acedo et al., 2010). The influence of peer pressure on criminal behavior (Burgess and Akers, 1966; Esiri, 2016; Harkins, Williams and Burger, 2017) is considered in a way reminiscent of a contagious disease; mathematically, the nonlinear terms in the compartments model complex interactions. Some papers on this type of models for crime evolution are given by Abbas, Tripathi and Neha (2017); González-Parra, Chen-Charpentier and Kojouharov (2018); Srivastav, Athithan and Ghosh (2020). Other disciplines employ the same approach, see Song et al. (2006); White and Comiskey (2007); Santonja et al. (2010) for drug or alcohol consumption, and Cervelló et al. (2014) for telecommunications, as good examples. Short reviews on this topic are published by Sooknanan and Comissiong (2017); Koss (2019). Recently, based on spatial epidemic models (van den Driessche, 2008; Wu, 2008; Schiesser, 2019), we extended this kind of compartmental model for crime to the case of spatial heterogeneity, by considering ordinary differential equations structured in patches and reaction-diffusion partial differential equations (Calatayud, Jornet and Mateu, 2025a).

These models contribute to the theory of mathematical criminology and the understanding of dynamical properties of criminality. However, because of the mechanisms involved in the models, it is certainly difficult to apply them when fitting to real data within a jurisdiction. The availability of adequate records, the lack of empirical values for model's parameters, or the divergence of the optimization procedure to calibrate the model's parameters (due to high dimensionality, stagnation at local minima, or unidentifiable values) are common issues when addressing the problem of data fitting. Therefore, the models are less useful for making specific predictions of crime events Kolokolnikov, Lloyd and Short (2019).

In the literature, works that aim to fit crime data with differential equations are more recent and scarce. Deterministic systems of ordinary differential equations were proposed by Lacey and Tsardakas (2016); Jane White et al. (2021). In Lacey and Tsardakas (2016), the authors studied serious and minor criminal events and built three coupled equations based on the attractiveness of the place. The monthly data on crime in Manchester were fitted by least-squares optimization. The authors highlighted that the inverse problem contained unidentifiable parameters. In Jane White et al. (2021), the authors used two coupled ordinary differential equations, based on population fluxes, to fit annual crime data in South Africa. The study region was divided into high- and low-conflicting areas. Bayesian inference was applied, with deterministic values reported. On the other hand, to accommodate fluctuations in crime time series and seek greater flexibility in modeling, stochastic differential equation models (Evans, 2012; Allen, 2007; Mao, 2007; Rackauckas, 2014) were investigated and calibrated in the

works by Cao et al. (2013); Calatayud, Jornet and Mateu (2025b; 2023a; 2023b). In the report by Cao et al. (2013), the authors considered daily burglary data from Los Angeles (California) and Houston (Texas), and trend time series were modeled (for each city independently) with a stochastic Lotka-Volterra system, with independent Brownian noises. Least-squares fitting and maximum-likelihood estimation were carried out. In Calatayud et al. (2025b), criminality in Spain was fitted with a compartmental system of three ordinary differential equations (susceptible individuals, offenders in liberty, and offenders imprisoned), considering social influence with a nonlinear term, deterministic estimates with least-squares minimization, and nonlinear regression for generating confidence bands. An analysis of the basic reproduction number, conceived as the force of criminality in the region, was carried out. In Calatayud et al. (2023a), the events of aggression, theft, and woman alarm in the city of Valencia (Spain), were modeled with the stochastic log-normal model, by correlating two Brownian motions. Multidimensional correlations, beyond two Brownian motions, or spatial effects were not studied. This motivated the contribution by Calatayud et al. (2023b), who proposed stochastic differential equations of spatio-temporal type to investigate time-independent correlations of criminality between the twenty-six zip codes of the city of Valencia. The corresponding stochastic log-normal models were related by correlating Brownian processes.

The present paper continues the research of our previous contribution (Calatayud et al., 2023b). In particular, we aim at:

- Building a stochastic partial differential equation model for spatially referenced crime data with a random-field intensity, based on stochastic log-normal models;
- Studying spatio-temporal correlations of crime between regions, with transformed stochastic log-normal models;
- Framing all pieces of our model in a Bayesian hierarchical structure that provides an adequate framework for inference and forecasting;
- Illustrating the results on the real crime data of Valencia, with inference of the model's parameters.

Concerning the first point, we provide an adequate framework to Calatayud et al. (2023b), based on stochastic partial differential equations. The first paper that applied stochastic partial differential equations to real-data fitting in a spatial framework was proposed by Duan, Gelfand and Sirmans (2009), where urban housing development was modeled through a logistic-growth intensity function and Gaussian-process disturbances with a Matérn spatial covariance function (Gelfand et al., 2010) (but with no crime processes involved). Bayesian inference was used for inverse parameter estimation (Banerjee, Carlin and Gelfand, 2014; Lesaffre and Lawson, 2012). Such a mechanistic form for the intensity function was due to the clear sigmoid growth of the data. In our proposal, we focus on crime data instead, for the first time, and adopt a stochastic log-normal model, based on the financial literature (Lamberton and Lapeyre, 2011; Voit, 2010). The correlation is alternatively set by relating Brownian motions in a discretized space. Parameter estimation is computationally tractable within the Bayesian paradigm.

Regarding the second point, we extend the constant correlations in Calatayud et al. (2023b) to time-varying stochastic correlations. Some papers in finance treated stochastic correlations (van Emmerich, 2006; Teng, Ehrhardt and Günther, 2016) (the contribution by van Emmerich (2006) seems to be the pioneer). Parameter inference was conducted by fitting a stationary density to the empirical density of the historical-correlation time series. Nonetheless, the model was not tested on historical asset prices or correlations, but it was employed for option pricing. In our context of crime, we estimate the parameters for the transformed log-normal model instead, by using the Bayesian approach; these are the third and fourth points of the research. We are not aware of previous contributions that combine such methodology and techniques, and we think that it could be a good starting point for the analysis of several spatially distributed types of data. Here we particularize to the field of mathematical and statistical criminology, due to the significant importance of the topic in society.

The plan of the paper is the following. The methodological development comes in Section 2, where we present and develop the mathematical model, describe the corresponding differential equations, and frame the model into a hierarchical structure embedded in a Bayesian paradigm. Then, Section 3 presents the Bayesian implementation and the numerical and statistical results for the crime data of Valencia. The paper ends with some final considerations in Section 4, where a discussion and possible future extensions are given.

2. Formulation of the stochastic model

This section is devoted essentially to the design of the model. In Section 2.1, we start with a stochastic partial differential equation for the intensity function, with a spatially correlated noise and a temporal log-normal evolution. For geostatistical and lattice data, the model becomes finite dimensional and reduces to a set of stochastic differential equations, coupled by correlated Wiener processes. In Section 2.2, we further extend our proposal with a stochastic model for time-varying correlations. Hence, the intensity is in this case doubly stochastic and is framed within a Bayesian hierarchical structure.

2.1. Stochastic partial differential equations and spatio-temporal stochastic differential equations

An ordinary differential equation subject to an initial condition takes the form (Murray, 2002)

$$u'(t) = f(t, u(t), \Theta), \quad u(0) = u_0, \tag{1}$$

where $u(t)$ is the state variable with derivative $u'(t)$, f is a function that defines the model, Θ is a set of real parameters, and u_0 is the initial condition. This mathematical formula may be of use to model phenomena described by smooth curves, that are generally viewed as limits of difference equations. Although smoothness is rarely encountered in nature, an ordinary differential equation may be a useful tool to fit to average dynamics, especially when fluctuations in data are of small magnitude.

When fluctuations cannot be ignored, randomness can be incorporated into (1) through an irregular stochastic process, often a Gaussian white noise $B'(t)$ in $\mathbb{R}$ (Allen, 2007; Mao, 2007; Smith, 2013). The notation $B'(t)$ stands for a generalized derivative, since the formal differentiation of a standard Brownian-motion process $B(t)$ (Wiener process) yields a Gaussian white noise. The system (1) is then modified to (Evans, 2012; Allen, 2007; Mao, 2007; Rackauckas, 2014)

$$u'(t) = f(t,u(t),\Xi) + g(t,u(t),\Xi)B'(t), \quad u(0) = u_0,$$

where Ξ is a set of real parameters larger than or equal to Θ, g is a function that represents the intensity of the noise B', u is the stochastic state variable, and u_0 is the deterministic initial condition. Rigorously, the equation is interpreted in integral form under the theory of Itô calculus; therefore, a differential notation using d is employed, to give

$$\mathrm{d}u(t) = f(t,u(t),\Xi)\mathrm{d}t + g(t,u(t),\Xi)\mathrm{d}B(t), \quad u(0) = u_0. \tag{2}$$

Physically, and for modeling purposes, one may view the differential d as a very small increment. This pragmatic interpretation, which will be used extensively throughout the paper, agrees with the approximation of Itô stochastic differential equations by stochastic difference equations of Euler type (Euler-Maruyama scheme on a finite time mesh), in contrast to the Stratonovich formulation (Braumann, 2007). The increment $\mathrm{d}B(t) = B'(t)\mathrm{d}t$ is an uncorrelated Gaussian process with zero mean and variance $\mathrm{d}t$.

For modeling financial data (Lamberton and Lapeyre, 2011); Voit, 2010), where a single time series characterized by fluctuations is considered, an important model is usually employed: the log-normal equation or geometric Brownian-motion process, given by

$$\mathrm{d}u(t) = \mu u(t)\mathrm{d}t + \sigma u(t)\mathrm{d}B(t). \tag{3}$$

In the previous notation, $\Theta = \mu$, $\Xi = (\mu,\sigma)$, $f(t,u(t),\Xi) = \mu u(t)$ and $g(t,u(t),\Xi) = \sigma u(t)$. Parameter $\mu \in \mathbb{R}$, called drift, captures the growth rate, and $\sigma > 0$, called volatility, captures the magnitude of the fluctuations and the uncertainty on the future values. Essentially, the customary deterministic model for the infinitesimal growth rate, $(u(t+\mathrm{d}t) - u(t))/u(t) = \mu\mathrm{d}t$, is extended to a random setting as

$$\frac{u(t+\mathrm{d}t) - u(t)}{u(t)} \sim \mathrm{N}(\mu\mathrm{d}t, \sigma^2\mathrm{d}t), \tag{4}$$

with mutually independent perturbations, where N is the normal distribution [1]. By Itô's lemma, which extends the classical chain-rule theorem for non-differentiable processes,

[1] The randomization in (4) is the most consistent model. Indeed, consider a general model

$$\frac{u(t+\mathrm{d}t) - u(t)}{u(t)} = \mu\mathrm{d}t + \sigma(\mathrm{d}t)^{\gamma/2}\zeta_t,$$

where $\gamma \geq 0$ is a constant and ζ_t is an uncorrelated process with mean zero and variance one:

- The Gaussian behavior for ζ_t corresponds to the maximum-entropy distribution (Dorini and Sam-

the solution to (3) is given by the stochastic process

$$u(t) = u_0 \mathrm{e}^{(\mu - \frac{1}{2}\sigma^2)t + \sigma B(t)}.$$

The expected value of $u(t)$, $\mathbb{E}[u(t)]$, is the solution to the deterministic counterpart $\mathrm{d}\mathbb{E}[u(t)] = \mu \mathbb{E}[u(t)]\mathrm{d}t$,

$$\mathbb{E}[u(t)] = u_0 \mathrm{e}^{\mu t}. \tag{5}$$

Hence (3) can be interpreted as "mean" + "residual", for the differential. Since the probability distribution of $u(t)$ is log-normal, it is possible to compute a probabilistic interval at level $1-\alpha$,

$$\left[u_0 \mathrm{e}^{(\mu - \frac{1}{2}\sigma^2)t - \sigma \cdot \sqrt{t} \cdot q_{\alpha/2}}, u_0 \mathrm{e}^{(\mu - \frac{1}{2}\sigma^2)t + \sigma \cdot \sqrt{t} \cdot q_{1-\alpha/2}} \right], \tag{6}$$

where q is the quantile function of a $\mathrm{N}(0,1)$. The paths of u are continuous, autocorrelated, and nowhere differentiable. If $u_0 > 0$, then the paths of u are positive.

This log-normal model is not restricted to financial data; it can indeed be used whenever the observed fluctuations match the dynamics of (3). In fact, for crime time series, our prototype is (3). Intuitively, and despite the limitations, the evolution of crime incidence can be captured by the log-normal model, where there is a rate for the growth of criminality and a volatility for the random fluctuations. Albeit simple, the exponential model, rooted in birth-death environments, reflects imitative and social criminality (Burgess and Akers, 1966; Esiri, 2016; Harkins et al., 2017).

Until now, spatial effects have been omitted. To extend (3) to space, we consider a fixed compact region $D \subseteq \mathbb{R}^2$. Two settings may be considered: geostatistical data, where fixed (non-random) locations $s_1, \ldots, s_n \in D$ are studied, or lattice data, where D is partitioned into fixed (non-random) disjoint regions $D_1, \ldots, D_n$. In the context of criminology, the set D could represent a city of interest, for example. Lattice models divide D, for instance, into districts or streets. Geostatistical models deal with points of reference in the city, where we can have measurements of the quantity of interest. Strictly, these spatial formulations are not the same, but points of reference could also define a partition $\tilde{D} = \{s_1, \ldots, s_n\} = \{s_1\} \cup \ldots \cup \{s_n\}$ or $\tilde{D} = U_1 \cup \ldots \cup U_n$, where each U_i is a neighborhood of the representative s_i.

---

paio, 2012) (i.e. it has the minimum amount of information). Given an unbiased error, it is reasonable to presume a symmetric shape for its probability density. The normal distribution is mathematically convenient and simple.

- For $\gamma > 2$, the model is actually deterministic: $u' = \mu u$. For $\gamma \leq 2$, denote $\mathrm{d}G(t) = (\mathrm{d}t)^{\gamma/2}\zeta_t$. The correlation function of the noise, $\frac{\mathrm{d}G(t)}{\mathrm{d}t} = \frac{\zeta_t}{(\mathrm{d}t)^{1-\gamma/2}}$, is $C(t,\tau) = \mathbb{E}[\frac{\mathrm{d}G(t)}{\mathrm{d}t}, \frac{\mathrm{d}G(\tau)}{\mathrm{d}\tau}]$. For $t \neq \tau$, $C(t,\tau)\mathrm{d}\tau = 0$, because $\mathbb{E}[\zeta_t \zeta_\tau] = \mathbb{E}[\zeta_t]\mathbb{E}[\zeta_\tau] = 0$. For $t = \tau$, $C(t,\tau)\mathrm{d}\tau = \frac{1}{(\mathrm{d}\tau)^{1-\gamma}}$. The correlation only has an adequate structure when $\gamma = 1$, since in such a case, $C(t,\tau)\mathrm{d}\tau$ acts as the Dirac measure $\mathrm{d}\delta_t(\tau)$ and $C(t,\tau)$ acts as the Dirac-delta generalized function $\delta(t-\tau)$. Then, $\mathrm{d}G(t) = \mathrm{d}B(t)$. For $\gamma < 1$, $C(t,\tau) = \infty$ in the sense of distributions. For $\gamma > 1$, $C(t,\tau) = 0$ in the sense of distributions.

The process of interest in this paper is the (surface) intensity of crimes, $\Lambda(s,t)$, which is stochastic in both space and time. Conceptually, it is the expected number of crimes per unit area at coordinate s on the temporal interval $(t-1,t]$, for $s \in D$ and $t \in [0,T]$. The time horizon $0 < T < \infty$ comes up because data are always collected within a bounded period. This surface intensity on unit-time intervals is the integral of the surface-time intensity $\Gamma(s,\tilde{t})$:

$$\Lambda(s,t) = \int_{t-1}^{t} \Gamma(s,\tilde{t})\mathrm{d}\tilde{t}.$$

The intensities measure the risk of crime, hence an increase in the intensity could indicate a need of police intervention.

Based on (3), a stochastic partial differential equation model for the random field Λ is given by

$$\frac{\partial \Lambda(s,t)}{\partial t} = \mu(s)\Lambda(s,t) + \sigma(s)\Lambda(s,t)\xi(s,t), \tag{7}$$

for each point s within D, where $\mu(s) \in \mathbb{R}$ and $\sigma(s) \in (0,\infty)$ are spatially heterogeneous non-random functions, and ξ is a certain unbiased spatio-temporal noise independent of Λ, without further specification for now. The notation ∂ means a partial derivative, with respect to t here. In (7), the temporal evolution is given by (3), whereas the spatial distribution is determined by the noise ξ, which exhibits a certain spatial correlation. The system is subject to an initial state $\Lambda(s,0) = \int_{-1}^{0} \Gamma(s,\tilde{t})\mathrm{d}\tilde{t}$, $s \in D$. No boundary conditions are needed.

It is interesting to observe that (7) is equivalent to a stochastic integro-difference equation, like that proposed by Zammit-Mangion et al. (2012). These models describe the conditional dependence between the spatial field at a future-time point and the field at the present-time point through an integral operator, which is typically assumed to be linear (Zammit-Mangion and Wikle, 2020). Given time instants $t_k < t_{k+1}$ with $\Delta t = t_{k+1} - t_k$, one has

$$\begin{aligned}\Lambda(s,t_{k+1}) &\approx \Lambda(s,t_k) + \mu(s)\Lambda(s,t_k)\Delta t + \sigma(s)\Lambda(s,t_k)\xi(s,t_k)\Delta t \\ &= \int_D K(s,\tilde{s})F(\tilde{s},\Lambda(\tilde{s},t_k))\mathrm{d}\tilde{s} + e_k(s),\end{aligned} \tag{8}$$

where $K(s,\tilde{s}) = \delta(s-\tilde{s})$ is a Dirac-delta kernel, $F(\tilde{s},\Lambda(\tilde{s},t_k)) = \Lambda(\tilde{s},t_k) + \mu(\tilde{s})\Lambda(\tilde{s},t_k)\Delta t$ distorts the field in the sedentary stage, and $e_k(s) = \sigma(s)\Lambda(s,t_k)\xi(s,t_k)\Delta t$ is a zero-mean, correlated spatial disturbance. The δ-kernel corresponds to negligible spatial interactions through the drift and the volatility; spatial dependencies arise from the stochastic noise e_k and its covariance structure. The stochastic integro-difference formulation will not be used in the subsequent development; we will base our arguments on the differential form (7) instead.

The stochastic partial differential equation (7) (infinite-dimensional model) can actually be reduced to a set of n stochastic differential equations (finite-dimensional model), by taking the spatial discretization into account. For geostatistical data with locations

$s_1,\ldots,s_n$,

$$\frac{\partial \Lambda(s_i,t)}{\partial t} = \mu(s_i)\Lambda(s_i,t) + \sigma(s_i)\Lambda(s_i,t)\xi(s_i,t). \tag{9}$$

For lattice data with grid cells $D_1,\ldots,D_n$, if $\Lambda(D_i,t) = \int_{D_i}\Lambda(s,t)\mathrm{d}s$ is the aggregated intensity, then

$$\frac{\partial \Lambda(D_i,t)}{\partial t} = \int_{D_i}\mu(s)\Lambda(s,t)\mathrm{d}s + \int_{D_i}\sigma(s)\Lambda(s,t)\xi(s,t)\mathrm{d}s.$$

If we assume that $\mu(s)$, $\sigma(s)$ and the noise $\xi(s,t)$ are spatially homogeneous within D_i, given by $\mu(D_i)$, $\sigma(D_i)$ and $\xi(D_i,t)$, respectively (i.e., intra-region homogeneity and between-regions heterogeneity), then

$$\int_{D_i}\mu(s)\Lambda(s,t)\mathrm{d}s = \mu(D_i)\int_{D_i}\Lambda(s,t)\mathrm{d}s = \mu(D_i)\Lambda(D_i,t)$$

and

$$\int_{D_i}\sigma(s)\Lambda(s,t)\xi(s,t)\mathrm{d}s = \sigma(D_i)\xi(D_i,t)\int_{D_i}\Lambda(s,t)\mathrm{d}s = \sigma(D_i)\Lambda(D_i,t)\xi(D_i,t).$$

In consequence, for lattice data, we have

$$\frac{\partial \Lambda(D_i,t)}{\partial t} = \mu(D_i)\Lambda(D_i,t) + \sigma(D_i)\Lambda(D_i,t)\xi(D_i,t). \tag{10}$$

Both equations (9) and (10) can be combined into a single model, with the appropriate interpretations for Λ_i, μ_i, σ_i and ξ_i [2]:

$$\frac{\partial \Lambda_i(t)}{\partial t} = \mu_i\Lambda_i(t) + \sigma_i\Lambda_i(t)\xi_i(t). \tag{11}$$

[2] We have checked that the stochastic partial differential equation model for the intensity $\Lambda(s,t)$ implies temporal stochastic differential equations for aggregated intensities. The converse is also true. Suppose that any aggregated intensity

$$\Lambda(E,t) = \int_E \Lambda(\tilde{s},t)\mathrm{d}\tilde{s}$$

satisfies the stochastic differential equation

$$\frac{\partial \Lambda(E,t)}{\partial t} = \mu(E)\Lambda(E,t) + \sigma(E)\Lambda(E,t)\xi(E,t),$$

for small neighborhoods $E \subseteq D$ of $s \in D$ with area $|E|$. Then the local model

$$\frac{\partial \Lambda(s,t)}{\partial t} = \mu(s)\Lambda(s,t) + \sigma(s)\Lambda(s,t)\xi(s,t)$$

is retrieved, since

$$\Lambda(s,t) = \lim_{|E|\to 0}\frac{1}{|E|}\int_E \Lambda(\tilde{s},t)\mathrm{d}\tilde{s},$$

$$\mu(s)\Lambda(s,t) = \lim_{|E|\to 0}\frac{1}{|E|}\int_E \mu(\tilde{s})\Lambda(\tilde{s},t)\mathrm{d}\tilde{s}$$

and

$$\sigma(s)\Lambda(s,t)\xi(s,t) = \lim_{|E|\to 0}\frac{1}{|E|}\int_E \sigma(\tilde{s})\Lambda(\tilde{s},t)\xi(\tilde{s},t)\mathrm{d}\tilde{s}.$$

Notice that for lattice data, $\Lambda_i(t)$ gives an absolute quantity, not a density: the number of crimes in region D_i, on $(t-1,t]$. Indeed, the spatial integration of an intensity gives the "mass" (usually, in the literature, researchers simply talk about the number of crimes in a region at t for easiness, to really refer to $(t-1,t]$). For geostatistical data, by contrast, $\Lambda_i(t)$ is still an intensity: the surface density of crimes at s_i, on $(t-1,t]$.

Equations in (11) are formulated in $\mathbb{R}$, instead of $\mathbb{R}^2$. Each $\xi_i(t)$ can be defined as a one-dimensional Gaussian white-noise process, given by the derivative of a one-dimensional Brownian motion, $B_i'(t)$. Therefore, (11) is rewritten in differential form as

$$\mathrm{d}\Lambda_i(t) = \mu_i\Lambda_i(t)\mathrm{d}t + \sigma_i\Lambda_i(t)\mathrm{d}B_i(t), \tag{12}$$

for $i = 1,\dots,n$. The closed-form solutions are

$$\Lambda_i(t) = \Lambda_i(0)\mathrm{e}^{(\mu_i-\frac{1}{2}\sigma_i^2)t+\sigma_i B_i(t)}. \tag{13}$$

The specification of a log-normal process is equivalent to the specification of a Gaussian process. We note that if $\Lambda_i(t) = \mathrm{e}^{z_i(t)}$, then (12) is equivalent to $\mathrm{d}z_i(t) = (\mu_i - \frac{1}{2}\sigma_i^2)\mathrm{d}t + \sigma_i\mathrm{d}B_i(t)$. This identity is related to the stochastic integro-difference equation proposed by Zammit-Mangion et al. (2012), not for the intensity itself with (8), but for the logarithm of the intensity. In Euler discrete form, $z_i(t_{k+1}) \approx z_i(t_k) + \varepsilon_i(t_k)$, where the mean of the Gaussian variable $\varepsilon_i(t_k)$ is $(\mu_i - \frac{1}{2}\sigma_i^2)\Delta t_k$ and the variance is $\sigma_i^2\Delta t_k$. The covariance structure between two indices i, j, under spatial heterogeneity, will be specified later. Note that this model is the stochastic integro-difference equation with Dirac-delta kernel K and identity map F.

The original two-dimensional noise $\xi(s,t)$ is given by

$$\xi(s,t) = \sum_{i=1}^{n} \mathcal{X}_{\{s_i\}}(s)B_i'(t), \quad s \in \{s_1,\dots,s_n\},$$

for geostatistical data, and by

$$\xi(s,t) = \sum_{i=1}^{n} \mathcal{X}_{D_i}(s)B_i'(t), \quad s \in \bigcup_{i=1}^{n} D_i = D,$$

for lattice data, where $\mathcal{X}_\cdot(\cdot)$ denotes the indicator function over a set. Under a higher level of abstraction, we note that bivariate noise is a particular form of the appealing expression

$$\xi(s,t) = \sum_i w_i(s)\eta_i(t),$$

where η_i are temporal noises with a certain interdependence over i, and w_i are spatial functions (weights) that distribute those temporal noises in space. Separation of variables decomposes the domain $D\times[0,T]$ into D and $[0,T]$ and simplifies the problem [3].

[3]The decomposition of Cartesian-product domains for random-field representations is a common tool in stochastic modeling. For instance, similar approaches are adopted for stochastic systems driven by parametric uncertainty, with Karhunen-Loève expansions or polynomial chaos expansions (Lord, Powell and Shardlow, 2014; Xiu, 2010).

When w_i are indicator functions, the space is then discretized and the stochastic partial differential equation becomes a set of stochastic differential equations.

In the notation of lattice data, but the same applies for geostatistical data, time-value properties of our ξ are: (a) $\xi(s,*)$ is a Gaussian process; (b) the expectation becomes $\mathbb{E}[\xi(s,t)] = \sum_{i=1}^{n} \chi_{D_i}(s)\mathbb{E}[B_i'(t)] = 0$; (c) and the covariance can be written as

$$\begin{aligned}\mathbb{C}\mathrm{ov}[\xi(s,t_1),\xi(s,t_2)] &= \sum_{i,j=1}^{n} \chi_{D_i}(s)\chi_{D_j}(s)\mathbb{C}\mathrm{ov}[B_i'(t_1),B_j'(t_2)] \\ &= \sum_{i=1}^{n} \chi_{D_i}(s)\mathbb{C}\mathrm{ov}[B_i'(t_1),B_i'(t_2)] \\ &= \delta(t_1-t_2)\sum_{i=1}^{n} \chi_{D_i}(s) = \delta(t_1-t_2),\end{aligned}$$

where δ is the Dirac delta function. That is, $\xi(s,*)$ is a Gaussian white-noise stochastic process.

To set the properties of $\xi(*,t)$, since $\xi(s,t)$ exhibited spatial association, the n Brownian motions are correlated. This feature couples the n equations (12). No coupling is due to the drift or the volatility, which does not seem to contribute to lower fitting or prediction accuracy in practice.

To our knowledge, such a decomposition of the spatio-temporal noise in stochastic partial differential equations has not been used previously.

When $n = 2$ or, in other terms, we work pairwise, one can define

$$\mathrm{d}B_2(t) = \rho_{1,2}(t)\mathrm{d}B_1(t) + \sqrt{1-\rho_{1,2}(t)^2}\mathrm{d}\tilde{B}_2(t),$$

where $\rho_{1,2}(t) \in [-1,1]$ is a function of t and $\tilde{B}_2(t)$ is an auxiliary Brownian motion that is independent of $B_1(t)$. This differential equation is rigorously interpreted by Itô integration,

$$B_2(t) = \int_0^t \rho_{1,2}(\tilde{t})\mathrm{d}B_1(\tilde{t}) + \int_0^t \sqrt{1-\rho_{1,2}(\tilde{t})^2}\mathrm{d}\tilde{B}_2(\tilde{t}).$$

By the bilinearity of the covariance operator, it is easy to see that the correlation becomes

$$\mathrm{corr}[\mathrm{d}B_1(t),\mathrm{d}B_2(t)] = \rho_{1,2}(t).$$

For $\Lambda_i(t)$, this property translates into

$$\mathrm{corr}[\mathrm{d}\Lambda_1(t),\mathrm{d}\Lambda_2(t)|\Lambda_1(t),\Lambda_2(t)] = \rho_{1,2}(t),$$

where the vertical bar denotes a conditional quantity. That is, $\rho_{1,2}$ measures how similar the (infinitesimal) variations of $\Lambda_1(t)$ and $\Lambda_2(t)$ are. In practice, this is a key quantity: if $\rho_{1,2} > 0$ is somewhat near 1, then a significant variation in crime incidence within region D_1 should make the public authorities and the police put their attention on D_2 also.

In the general case, the theoretical construction of n correlated Brownian motions is as follows. Given n independent auxiliary Brownian processes $\tilde{B}_1(t) = B_1(t)$, $\tilde{B}_2(t), \ldots, \tilde{B}_n(t)$, define

$$\begin{pmatrix} \mathrm{d}B_1(t) \\ \vdots \\ \mathrm{d}B_n(t) \end{pmatrix} = \mathrm{L}(t) \begin{pmatrix} \mathrm{d}\tilde{B}_1(t) \\ \vdots \\ \mathrm{d}\tilde{B}_n(t) \end{pmatrix},$$

where $\mathrm{L}(t)$ is a lower-triangular matrix with transpose $\mathrm{L}^\top(t)$ and $\mathrm{A}(t) = (\rho_{i,j}(t))_{i,j} = \mathrm{L}(t)\mathrm{L}^\top(t)$ is the Cholesky decomposition for the symmetric and positive definite matrix $\mathrm{A}(t)$ of correlation functions. Indeed, the relations

$$\mathrm{corr}[\mathrm{d}B_i(t), \mathrm{d}B_j(t)] = \rho_{i,j}(t) \tag{14}$$

and

$$\mathrm{corr}[\mathrm{d}\Lambda_i(t), \mathrm{d}\Lambda_j(t) | \Lambda_i(t), \Lambda_j(t)] = \rho_{i,j}(t)$$

hold. Also, for (14), notice that

$$\Sigma_{\mathrm{d}B} = \mathrm{L}\Sigma_{\mathrm{d}\tilde{B}}\mathrm{L}^\top = \mathrm{d}t\mathrm{L}\mathrm{L}^\top = \mathrm{d}t\,\mathrm{A},$$

where Σ denotes the covariance matrix, $B = (B_1, \ldots, B_n)^\top$ and $\tilde{B} = (\tilde{B}_1, \ldots, \tilde{B}_n)^\top$.

As a consequence, if $s \in D_k$ (or $s = s_k$) and $r \in D_l$ (or $r = s_l$), then

$$\begin{aligned} \mathbb{C}\mathrm{ov}[\xi(s,t)\mathrm{d}t, \xi(r,t)\mathrm{d}t] &= \sum_{i,j=1}^{n} \mathcal{X}_{D_i}(s)\mathcal{X}_{D_j}(r)\mathbb{C}\mathrm{ov}[B_i'(t)\mathrm{d}t, B_j'(t)\mathrm{d}t] \\ &= \sum_{i,j=1}^{n} \mathcal{X}_{D_i}(s)\mathcal{X}_{D_j}(r)\mathbb{C}\mathrm{ov}[\mathrm{d}B_i(t), \mathrm{d}B_j(t)] \\ &= \sum_{i,j=1}^{n} \mathcal{X}_{D_i}(s)\mathcal{X}_{D_j}(r)\rho_{i,j}(t)\mathrm{d}t = \rho_{k,l}(t)\mathrm{d}t \end{aligned}$$

and

$$\mathrm{corr}[\xi(s,t)\mathrm{d}t, \xi(r,t)\mathrm{d}t] = \rho_{k,l}(t).$$

This is the spatial structure of the noise term in the stochastic partial differential equation. Then, the spatial structure of the intensity becomes

$$\begin{aligned} \mathbb{C}\mathrm{ov}[\mathrm{d}\Lambda(s,t), \mathrm{d}\Lambda(r,t) | \Lambda(s,t), \Lambda(r,t)] &= \sigma(s)\sigma(r)\Lambda(s,t)\Lambda(r,t)\mathbb{C}\mathrm{ov}[\xi(s,t)\mathrm{d}t, \xi(r,t)\mathrm{d}t] \\ &= \sigma(s)\sigma(r)\Lambda(s,t)\Lambda(r,t)\rho_{k,l}(t)\mathrm{d}t \end{aligned}$$

implying that

$$\mathrm{corr}[\mathrm{d}\Lambda(s,t), \mathrm{d}\Lambda(r,t) | \Lambda(s,t), \Lambda(r,t)] = \rho_{k,l}(t),$$

where the differential is taken with respect to t. At different spatial locations, the infinitesimal increments of the intensity over time are dependent via such a ρ-function.

2.2. Stochastic correlations

The most basic spatial correlation model assumes that $\rho_{i,j}(t) \equiv \rho_{i,j}$ is time independent (Calatayud et al., 2023b). This is a good option from an "averaged" point of view, when one uses a single quantity to summarize the relationship between the two historical time series. Moreover, a confidence interval for the correlation gives a region where the (constant) quantity lies with high confidence (Davison and Hinkley, 1997). However, in general, the correlation changes with time. This fact has been observed for data in finance (van Emmerich, 2006; Teng et al., 2016), and for crime dynamics the same seems to occur. It is interesting to notice that, in terms of time-series modeling, spatial dependencies between regions are somewhat similar to assets' dependencies between companies in the same financial market; hence the mathematical connection between spatial statistics and finance. After selecting a time lag $\tau > 0$, sample correlations of the log returns can be computed at blocks $[t-\tau, t]$ (moving-window technique), and it is observed that a noisy time series describes the dynamics of the empirical correlation; when τ is lower, the fluctuations are higher.

Thus, for completeness we better consider the correlation $\rho_{i,j}(t)$ as a stochastic process. Since it must belong to the interval $[-1,1]$, an appropriate transformation is used. In particular, making use of the hyperbolic tangent and the logarithm functions, we define

$$\rho_{i,j}(t) = \tanh(\log(y_{i,j}(t))) \in [-1,1], \tag{15}$$

where $y_{i,j}(t)$ is described by a log-normal model, as in (3), and it is given by

$$\mathrm{d}y_{i,j}(t) = \mu_{i,j} y_{i,j}(t)\mathrm{d}t + \sigma_{i,j} y_{i,j}(t)\mathrm{d}W_{i,j}(t), \tag{16}$$

where $W_{i,j}(t)$ are independent Brownian motions (hence the correlation processes operate independently), $\mu_{i,j} \in \mathbb{R}$, and $\sigma_{i,j} > 0$, for $i, j = 1, \ldots, n$. Notice that the hyperbolic tangent function has the image in $(-1,1)$, so $\rho_{i,j}$ is well-defined.

After combining (12), (14), (15) and (16), the final model we propose in this paper becomes hierarchical, and takes the form

$$\begin{cases} \mathrm{d}\Lambda_i(t) = \mu_i \Lambda_i(t)\mathrm{d}t + \sigma_i \Lambda_i(t)\mathrm{d}B_i(t), \quad i = 1, \ldots, n, \\ \mathrm{corr}[\mathrm{d}B_i(t), \mathrm{d}B_j(t)|W_{i,j}] = \mathrm{corr}[\mathrm{d}\Lambda_i(t), \mathrm{d}\Lambda_j(t)|\Lambda_i(t), \Lambda_j(t), W_{i,j}] = \rho_{i,j}(t), \\ \rho_{i,j}(t) = \tanh\left(\log\left(y_{i,j}(t)\right)\right), \\ \mathrm{d}y_{i,j}(t) = \mu_{i,j} y_{i,j}(t)\mathrm{d}t + \sigma_{i,j} y_{i,j}(t)\mathrm{d}W_{i,j}(t), \quad i, j = 1, \ldots, n. \end{cases} \tag{17}$$

Compared to (14), the second line in the hierarchy now conditions on $W_{i,j}$, which is the source of randomness for $\rho_{i,j}$; each path of $W_{i,j}$ defines a time function $\rho_{i,j}$. A loose schematic view on (17) is the following

$$\mathrm{data}_i \sim [\Lambda_i | B_1, \ldots, B_n; \mu_i, \sigma_i] \times [B_i, B_j | \rho_{i,j}] \times [\rho_{i,j} | W_{i,j}; \mu_{i,j}, \sigma_{i,j}] \times [W_{i,j}],$$

where the vertical bar conditions on random quantities, the semicolon conditions on (non-random) parameters, and the products multiply probability laws. Recall that the

subscripts i, j refer to spatial conditions: for geostatistical data, i refers to the fixed spatial location $s_i \in D$, whereas for lattice data, i refers to the region $D_i \subseteq D$. On the other hand, by Itô's lemma, one may derive a stochastic differential equation for $\rho_{i,j}(t)$ after some computations, without $y_{i,j}(t)$, but it will not be used for calibration purposes.

Mechanistically, model (17) is interpreted in the context of crime dynamics as follows. Omitting the subscript i for simplicity, the deterministic part of the first equation is $\Lambda'(t) = \mu\Lambda(t)$, for the region i. By relating number of crimes with number of criminals as $\Lambda(t) = \alpha x(t)$, where $\alpha > 0$ is a proportionality constant that reflects the average quantity of committed crimes per criminal in region i (Jane White et al., 2021), a model for the number of criminals is $x'(t) = \mu x(t)$. This equation takes into account: the social influence of criminality (Burgess and Akers, 1966; Esiri, 2016; Harkins et al., 2017), with an inflow $\mu_{\text{in}}x(t)$ that generates new offenders, and the cessation of criminal activity, with an outflow $\mu_{\text{out}}x(t)$, that gives rise to the continuous evolution

$$x(t+\mathrm{d}t) = x(t) + (\mu_{\text{in}} - \mu_{\text{out}})x(t)\mathrm{d}t.$$

Parameter μ, rooted in birth-death processes of populations, is thus interpretable as the balance between social influence and cessation of criminality, and it determines x when viewed as a deterministic function. However, there are uncertainties associated to crime evolution, hence the deterministic formulation for the relative change

$$\frac{x(t+\mathrm{d}t) - x(t)}{x(t)} = \mu\mathrm{d}t$$

is normally perturbed as

$$\frac{x(t+\mathrm{d}t) - x(t)}{x(t)} \sim \mathrm{N}(\mu\mathrm{d}t, \sigma^2\mathrm{d}t).$$

Parameter σ represents the volatility of the dynamics, for each region i. The mathematical formalization of this last equation gives rise to the stochastic differential equation of the Itô type. Now, the distinct regions are not independent and are related via correlated noises. For a higher fitting potential, correlations are also taken to be stochastic, but we employ a phenomenological formulation for them because we have no prior sociological information about the evolution. The use of the auxiliary functions tanh and log are mathematical artifacts to ensure that the correlation processes remain within $(-1, 1)$.

3. Crime data dynamics from calls to 112-emergency phone

This section presents the application of our methodological approach to the analysis of crime dynamics in the city of Valencia, Spain. In particular, we develop the framework and strategy for Bayesian inference to fit our stochastic differential equations to our crime data. In Section 3.1, we present the data and provide some notations. Then Section 3.2 outlines the Bayesian strategy for statistical inference, and presents the results: posterior distributions, fitting to the historical data, and predictions.

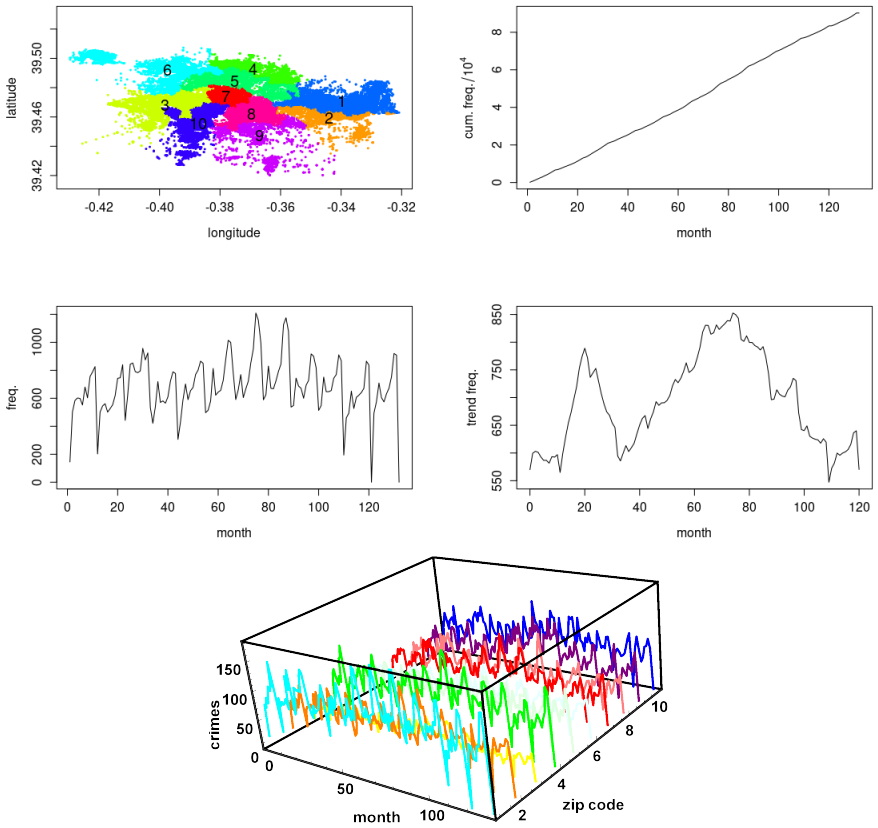

Figure 1. First panel*: Locations of crime events in Valencia partitioned in colors per each one of the* 10 *postal codes, for the period from January* 2010 *to December* 2020*. Then, each zip code has a time series of monthly crime counts.* Second panel*: Cumulative number of crime incidents in Valencia.* Third panel*: Frequency of crime events in Valencia.* Fourth panel*: Trend frequency (annual moving average) of crime events in Valencia.* Fifth panel*: 3D box showing the ten temporal series for each postal code.*

3.1. Crime-related setting: methodological setup

We work in the context of the recent papers by Calatayud et al. (2023a,b). The data consist of 90247 street crime incidents communicated to the 112-emergency phone in the Mediterranean city of Valencia, Spain, from January 2010 to December 2020. Essentially, these correspond to violent, smooth robberies and thefts in the streets. Valencia has around 800,000 inhabitants, ranked third in Spain, and it is the capital of the Valencian region; as a large and populated city, crime patterns stand as an important societal issue. Monthly counts of crime (132 measurements along 11 years) are available, positioned here in 10 regions of the city of Valencia based on their postal or zip codes.

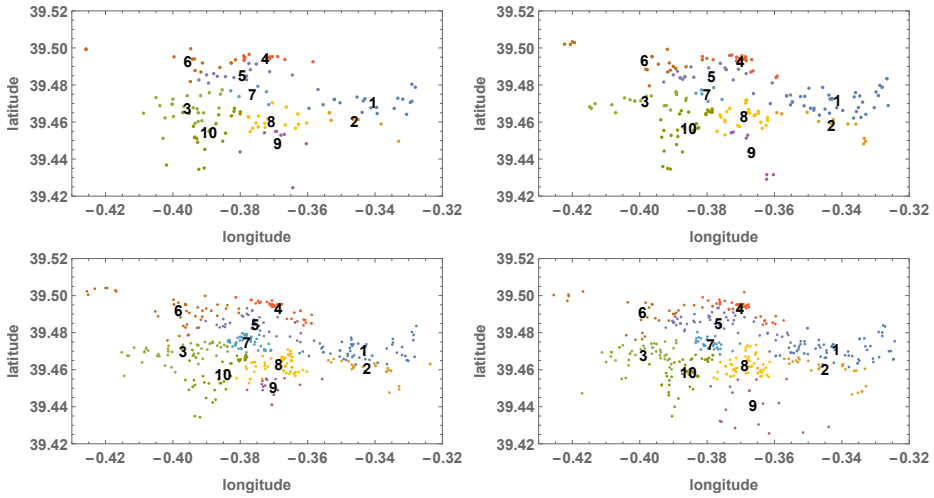

Figure 2. *Monthly spatial distribution of crime for the first four months of the data for all years: January (top-left), February (top-right), March (bottom-left), and April (bottom-right).*

The city is represented in the plane by the set D. The regional partition into 10 zip codes provides the context for lattice-data analysis. We have districts $D_i \subseteq D$, for $i = 1,\ldots,n$, where $n = 10$ and $D = \cup_{i=1}^{n} D_i$. The random-field intensity $\Lambda(s,t)$, defined as the expected number of crimes per unit area at s during month t (temporal interval $(t-1,t]$), is used in terms of aggregated intensity ("mass"): $\Lambda_i(t) = \Lambda(D_i,t) = \int_{D_i} \Lambda(s,t)\mathrm{d}s$. The integration of the density gives $\Lambda_i(t)$, representing the number of crimes in district D_i, at month t (temporal interval $(t-1,t]$). We assume a proportional relationship in terms of average number of crimes committed per criminal (Jane White et al., 2021), per zip code: incidents $= \alpha \times$ offenders, $\alpha > 0$; we then focus on the number of crimes. The time horizon is $T = 131$, and the time interval $[0,131]$ is partitioned into $t_1 < \ldots < t_m$, $m = 132$, with $t_{k+1} - t_k = \Delta t = 1$. The empirical data on street crimes are represented by λ_{i,t_k}. There are $n = 10$ time series of interest $\{\lambda_{i,t_k}\}_{k=1}^{m}$, for $i = 1,\ldots,n$. The raw time series are very noisy and Itô processes do not fit well, so an annual moving average is applied from the beginning to remove seasonality and outliers, and to accommodate geometric Brownian motions to trends, giving $\lambda_{i,t_k} \leftarrow \frac{1}{12}\sum_{\ell=0}^{11} \lambda_{i,t_{k-\ell}}$ and $m = 121$. This type of data processing was suggested by Cao et al. (2013). The new and smoother trend time series $\{\lambda_{i,t_k}\}_{k=12}^{m}$ are those used to fit our model. Indeed, while the raw time series are similar to a white noise due to the abrupt variability observed, the corresponding trend time series are better described as an Itô process (Cao et al., 2013; Calatayud et al., 2023b). The locations of crimes in Valencia are depicted in Figure 1, indicating each of the 10 zip codes that we are considering here. The bottom panel of Figure 1 shows basic information on the temporal evolution of the incidents, based on the cumulative frequency in Valencia. The third picture depicts the overall crime frequency in the city, in form of a time series. In the fourth panel, the time series is smoothed with the annual moving average. The fifth panel displays a $3D$ box containing the ten temporal series for

each postal code, enabling a simultaneous view of crime across both spatial and temporal dimensions.

To further investigate the temporal and spatial dynamics, Figure 2 presents the spatial distribution of crimes for the first four months for all years. Finally, Figure 3 shows the spatial distribution of crimes for the first four years. Each map represents a single year, allowing to identify and compare long-term spatial patterns.

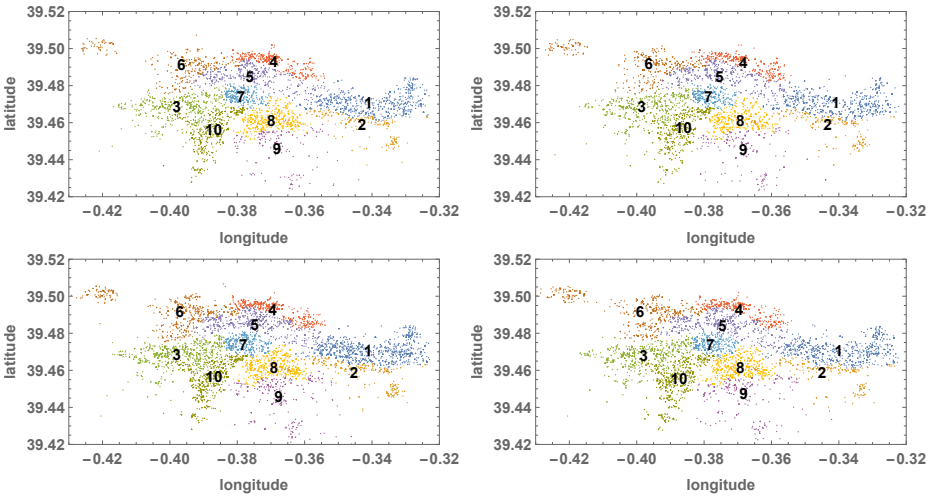

Figure 3. *Annual spatial distribution of crime for the first four years: Year 1 (top-left), Year 2 (top-right), Year 3 (bottom-left), and Year 4 (bottom-right).*

These figures collectively contribute to a more comprehensive understanding of the spatial distribution of crime in Valencia and its temporal evolution that will be interpreted through our stochastic model.

3.2. Bayesian inference for hierarchical model estimation

We assume discrete data labeled as $\{\lambda_{i,t_k}\}$, with $k = 1,\ldots,m$ and $i = 1,\ldots,n$, where subscript i refers to space, t_k represents an instant of time, and the time lags $\Delta t = t_{k+1} - t_k = 1$ are constant. These records are regarded as realizations of our hierarchical model of the Itô type.

We used the software Mathematica (Wolfram Research, 2020), for some preliminary computations with the data, and R (R Core Team, 2023) for the complete self-implementation of the Bayesian inference framework.

Given the data at the i-th location, $\lambda_{i,t_1},\ldots,\lambda_{i,t_m}$, we use a Bayesian approach following a Markov Chain Monte Carlo (MCMC) method for parameter estimation. Specifically, we use Gibbs sampling for the one-dimensional parameters and slice sampling for the multivariate settings (i.e., $Z_{i,t}$ and $W_{i,j}$, as coming next) (Brooks et al., 2011; Neal, 2003). Furthermore, we note that Λ_i, as defined in (17), may show negative intensity values when the stochastic differential equation is discretized with an Euler-Maruyama

scheme of the form

$$\begin{aligned}\Lambda_i(t_{k+1}) &= \Lambda_i(t_k) + \mu_i\Lambda_i(t_k) + \sigma_i\Lambda_i(t_k)(B_i(t_{k+1}) - B_i(t_k)) \\ &\sim \mathrm{N}\left(\Lambda_i(t_k) + \mu_i\Lambda_i(t_k), \sigma_i^2\Lambda_i(t_k)^2\right),\end{aligned}$$

$$\mathrm{support}(\Lambda_i(t_{k+1})) \in (-\infty, \infty).$$

which is not realistic. Therefore, we use a logarithmic transformation by taking advantage of Itô's lemma, see (13), and we propose the following hierarchical model, originated from (17):

$$\lambda_i(t)|\Lambda_i(t) \sim \mathrm{Po}\left(\Lambda_i(t)\right) \tag{18}$$

with Po() standing for a counting Poisson distribution, and

$$\log\left(\Lambda_i(t+1)\right) = \log\left(\Lambda_i(t)\right) + \left(\mu_i - \frac{1}{2}\sigma_i^2\right) + \sigma_i Z_{i,t}, \tag{19}$$

with

$$(Z_{1,A_l}, \ldots, Z_{n,A_l}) \sim \mathrm{N}_n(0, \Sigma_{A_l}), \quad \Sigma_{A_l} = (\rho_{i,j,A_l})_{1\le i,j\le n}, \tag{20}$$

where A_l are chosen subsets of the complete temporal interval. In our case, we divided the total temporal region into $L = 10$ subintervals A_l, each corresponding to one year, so $l = 1, \ldots, L = 10$. In addition,

$$\begin{aligned}\rho_{i,i,A_l} &= 1, \qquad \forall\, 1 \le i \le n, \\ \rho_{i,j,A_l} &= \tanh(\log(y_{i,j}(A_l))),\end{aligned} \tag{21}$$

where

$$\log\left(y_{i,j}(A_{l+1})\right) = \log\left(y_{i,j}(A_l)\right) + \left(\mu_{i,j} - \frac{1}{2}\sigma_{i,j}^2\right) + \sigma_{i,j}W_{i,j}. \tag{22}$$

Equation (18), added to (17) in practice, describes events that appear randomly and independently on a continuous space, characterized by the occurrence rate (the "intensity"), which is the expected number of cases per unit area.

We stress that in this model with n subregions (postal codes) and L partitions A_l of the overall temporal region, the number of parameters amounts to $3\binom{L}{2} + 3n$ plus those of the slice sampling over Z and W. In our case, both L and n are 10, resulting in a total of 165 parameters. This large number over-parametrizes our model, posing difficulties in the MCMC convergence and enlarging computational times unnecessarily. In consequence, we consider two stages. We first assume that the $Z_{i,t}$ are independent with respect to space, meaning that we estimate each time series separately and take the posterior median of each parameter as the correct value. Once this estimation is done, we then consider the spatial dependence of the $Z_{i,t}$ as defined in equation (20).

We also emphasize the fact that, although the conception of model (17) is of continuous type, the Bayesian implementation requires a discretization. Essentially, stochastic

processes become finite-dimensional random vectors and stochastic differential equations become random difference equations. Euler-type discretizations are broadly employed in stochastic computing; they are the simplest convergent methods, with strong convergence of order $1/2$. Milstein scheme, of higher convergence order in general, is not needed for parameter estimation, at least in our context.

On the other hand, we notice that Calatayud et al. (2023b) assumed constant correlations in time, and inference proceeded with the method of moments due to the simplicity of the solution in (13). In Calatayud et al. (2023b), correlations were estimated pairwise, by the underlying multivariate-Gaussian structure. Here, by contrast, the distributions of the parameters are estimated jointly. The downside is that, because of the higher complexity, blocks of time A_l and of space need to be considered.

The MCMC process was conducted over 20,000 iterations, with the initial 5,000 used as a "burn-in" period to ensure convergence. Samples of the parameters were taken at intervals of 50 iterations (thinning to reduce autocorrelation), resulting in a total of 300 values that provide an approximation of the posterior marginal distribution of all parameters involved in the process. We considered non-informative prior distributions for the spatial trends $\mu_i, \mu_{i,j}$ (a Gaussian distribution with a large variance), and informative priors for $\sigma_{i,j}, \Lambda_i(0)$, as follows:

$$\begin{aligned}
\mu_i &\sim \mathrm{N}(0,10),\\
\sigma_i &\sim \mathrm{U}(0.01,0.09),\\
\Lambda_i(0) &\sim \mathrm{Ga}(10,1),\\
\mu_{i,j} &\sim \mathrm{N}(0,10),\\
\sigma_{i,j} &\sim \mathrm{Ga}(1,1),\\
y_{i,j}(A_0) &\sim \mathrm{Ga}(1,1),\\
W_{i,j} &\sim \mathrm{N}(0,1),
\end{aligned}$$

for all postal codes, where N, Ga and U refer to normal, gamma and uniform random variables, respectively. The corresponding informative priors were selected empirically, to reduce the computational time and better adjust the domains of the outputs. We indeed stress that we have consistently employed non-informative priors throughout the modeling process. However, as the number of iterations in the Monte Carlo process increased, we observed that keeping excessively broad parameter domains led to a significant increase in computational cost, and posed challenges to model convergence. In order to improve both the stability and efficiency of the model, we empirically refined the parameter domains as the iterations progressed. This refinement was solely based on the behavior of the model during the initial stages of convergence and was not derived from any subset of the data used later for model fitting or predictive performance evaluation. This approach helped ensure computational feasibility while avoiding any risk of information leakage from the data used in subsequent phases of the analysis.

Figure 4 shows a comparison between the expected number of cases obtained from the posterior median of each parameter and the number of current cases, for some selected cases. Recall that we are working with trend time series. The estimates are quite accurate and capture well the temporal behavior of the observed number of cases over time.

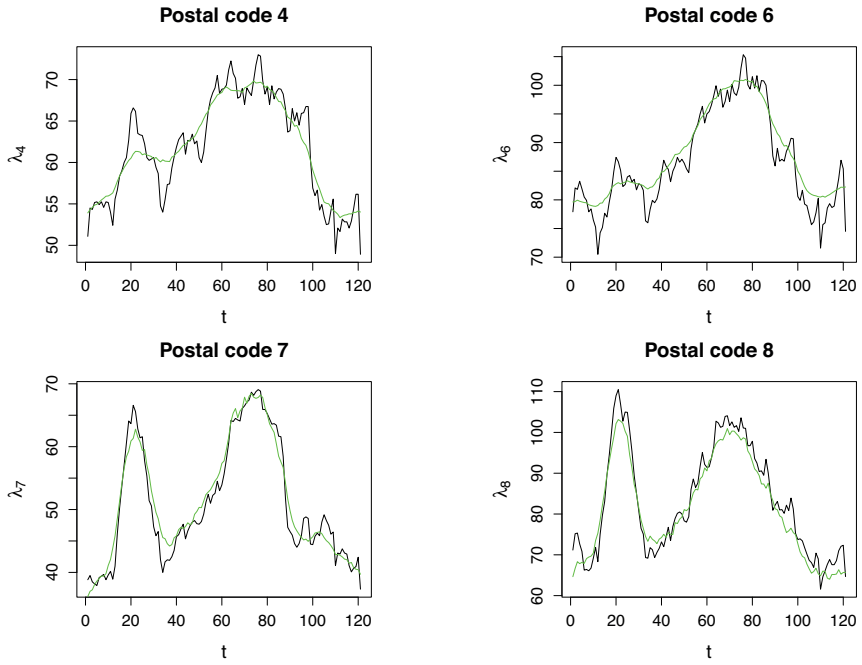

Figure 4. *Comparison between the estimate of cases (green line) and the number of real cases (black line), for postal codes 3, 6, 7 and 8.*

Our model also produces estimates of the correlations between any two postal codes, i.e., up to 45 pairs. To save space, here we focus on the pair $(1,5)$ and show the posterior densities of the parameters associated with the correlation between these two postal codes. Figure 5 depicts such posterior densities, specifically $\mu_{1,5}$, $\sigma_{1,5}$, $y_{1,5}(A_0)$ and $\rho_{1,5,A_0}$, noting that when using non-informative priors, the likelihood is very informative and drives the posterior distribution to the right place, with a shape similar to a bell. We show in the bottom right corner the posterior density of the correlation between postal codes 1 and 5. We obtained this sample by calculating the correlation value with equation (21) employing the corresponding parameter values in each iteration. From this sample, we obtained the estimate for each of the correlations by choosing the posterior median and the credibility intervals which we show in Figure 6, where we observe the behavior of the correlation across the subintervals A_l. We note the temporal variation of such correlations that our model has captured considering them stochastic processes themselves. The uncertainty of the estimates is given by the credible intervals. In general, the zero-correlation value does not lie within the credible region, suggesting a spatial effect in criminal behavior.

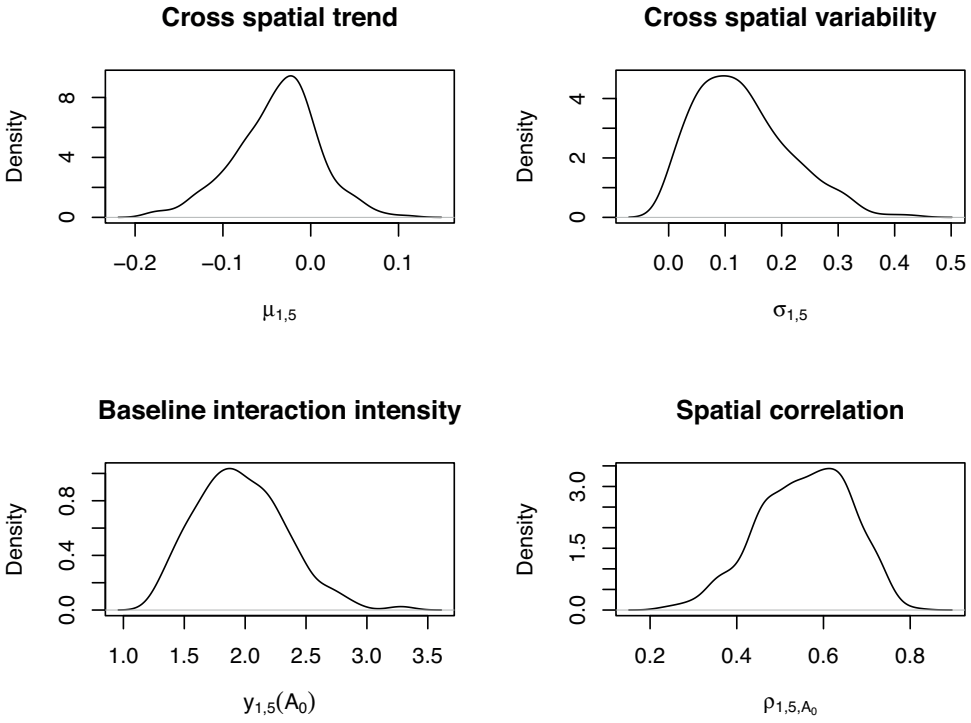

Figure 5. *Posterior densities of the parameters associated with the correlation between postal codes* 1 *and* 5*, as well as the posterior distribution of the correlation in the first subinterval* A_1*. The parameters are defined in equations* (21) *and* (22)*.*

Finally, we assess the accuracy of future predictions on the number of crime cases. We proceed in two ways: complete extrapolation and month-by-month extrapolation. For the former way, we use the estimated model for the first 9 years and predict the expected number for the last year (12 temporal instances), comparing them with the current data. Figure 7 depicts such predictions for the last year in four different postal codes, observing that the black line associated with the number of observed cases is almost always within the 0.95 interval. This is an indication that our proposed model has been accurately estimated and fits the crime data in Valencia well. The credible bands tend to widen because we are extrapolating for a long time window, along a complete year, hence there is a large uncertainty. Mathematically, the behavior is a consequence of the linearly increasing variance of the Brownian motion.

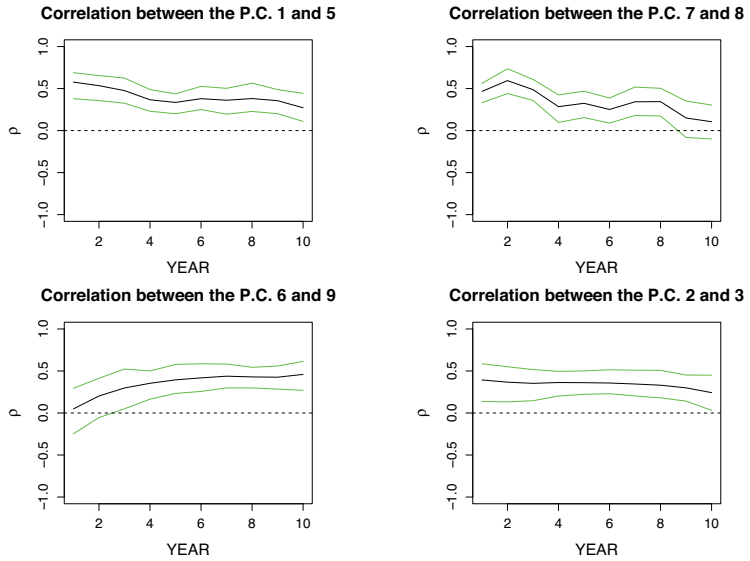

Figure 6. *Estimated pairwise correlations between four selected pairs of different postal codes and their corresponding* 0.95 *credibility intervals (green lines).*

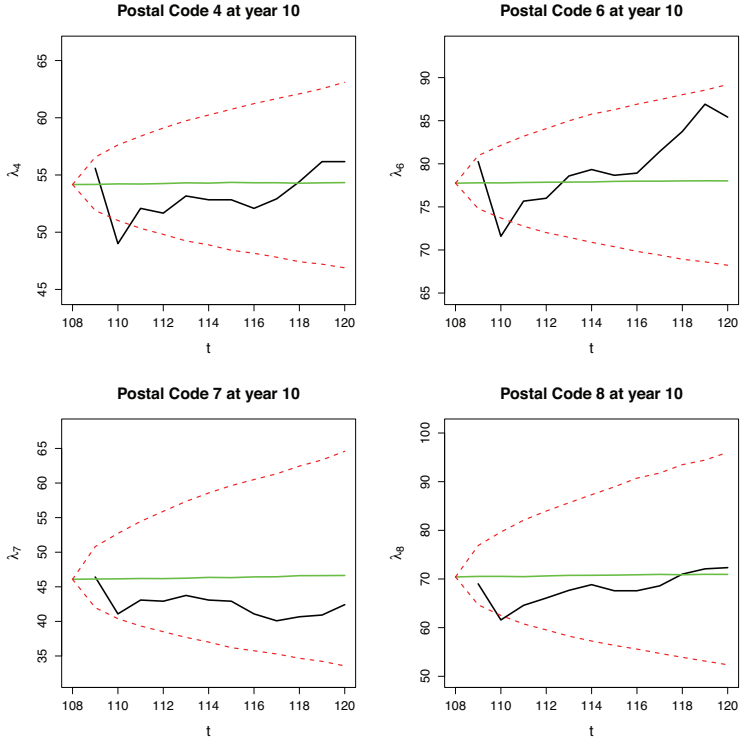

Figure 7. *Complete-one-year predictions (green line) and their credibility intervals (red line) for the last 12 months in 4 different areas, along with the observed cases (black line).*

However, if one forecasts month by month, the magnitude of the credible intervals is that of one time step. This is the case of the second approach: we predict the instant "9 years + ith month" from the data at instants $\leq$ "9 years + $(i-1)$th month", for $i \geq 1$. Figure 8 displays the new forecasts for the last year, this time for all zip codes for generality. The mean values and the credible intervals now move with the data, exhibiting narrower uncertainty. We remark that the last year corresponds to a very complicated scenario, coinciding with the COVID-19 pandemic and the lockdown in Spain (BOE, 2020; Wu et al., 2020). The second black data value of the pictures, that seems to be an outlier, is influenced by March 2020, where the restrictions on movement were imposed and reported criminality suddenly decreased. The fact that our model is successful for such a year makes us think that the proposed methodology is certainly useful for delineating crime dynamics in the short term. We note that the green and red lines advance in parallel, but it seems that there is a one-month lag with the black curve. Such a behavior is linked with the essence of prediction. For example, if there is a decrease of data from $t-1$ to t, this affects the parameter estimation for the times up to t, hence there will probably be a decrease in the prediction on $(t, t+1]$. If data increase from t to $t+1$, then the predicted values will likely rise on $(t+1, t+2]$.

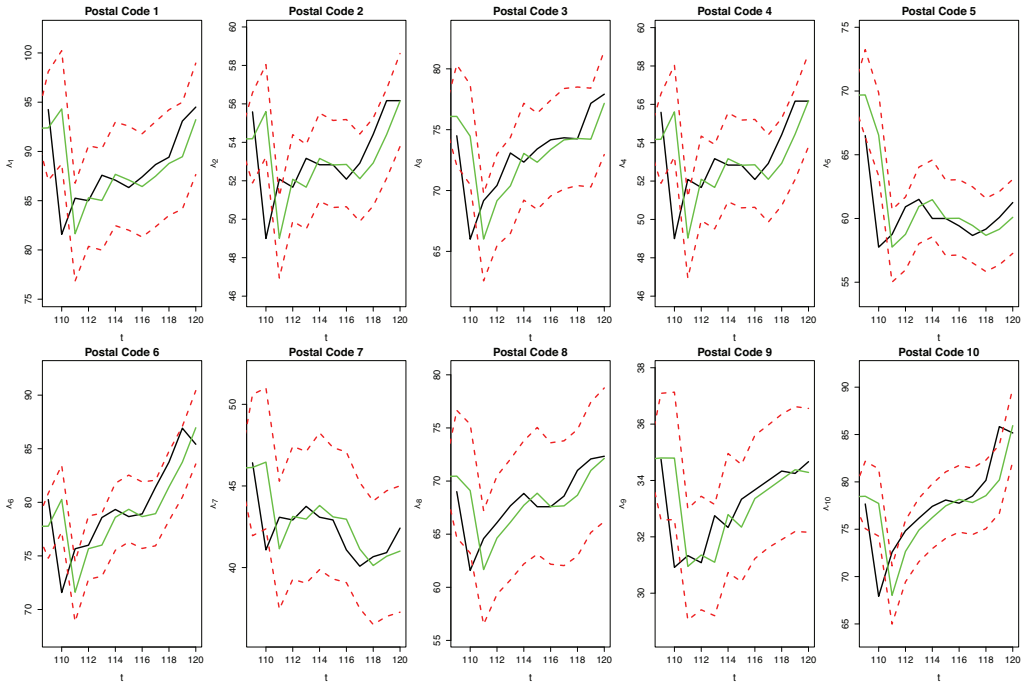

Figure 8. *Month-by-month predictions (green line) and their credibility intervals (red line) for the last 12 months (for year 10), along with the observed cases (black line).*

3.3. Comparison with alternative simpler models

To strengthen our approach, we offer a brief comparison among our model (HDEM), an autoregressive AR(1) structure, and a random forest (RF), as a widely used machine learning method.

Figure 9 presents comparatively credibility intervals (CI) for AR(1), random forest (RF), and HDEM for the case of a one-year prediction for some selected postal codes (3, 4, 6, 10) and year 10. We only show these four cases as an example to reinforce our arguments in favor of our model, since all the cases show similar behavior. We note that in terms of CIs, our HDEM and AR are quite close, suggesting similar levels of precision and reliability in the predicted outcomes. However, the CI of the RF seems to go away from the true values, indicating wrong predictions in all cases. We also considered the case of AR(p) for various orders, and the results were even worse compared to AR(1).

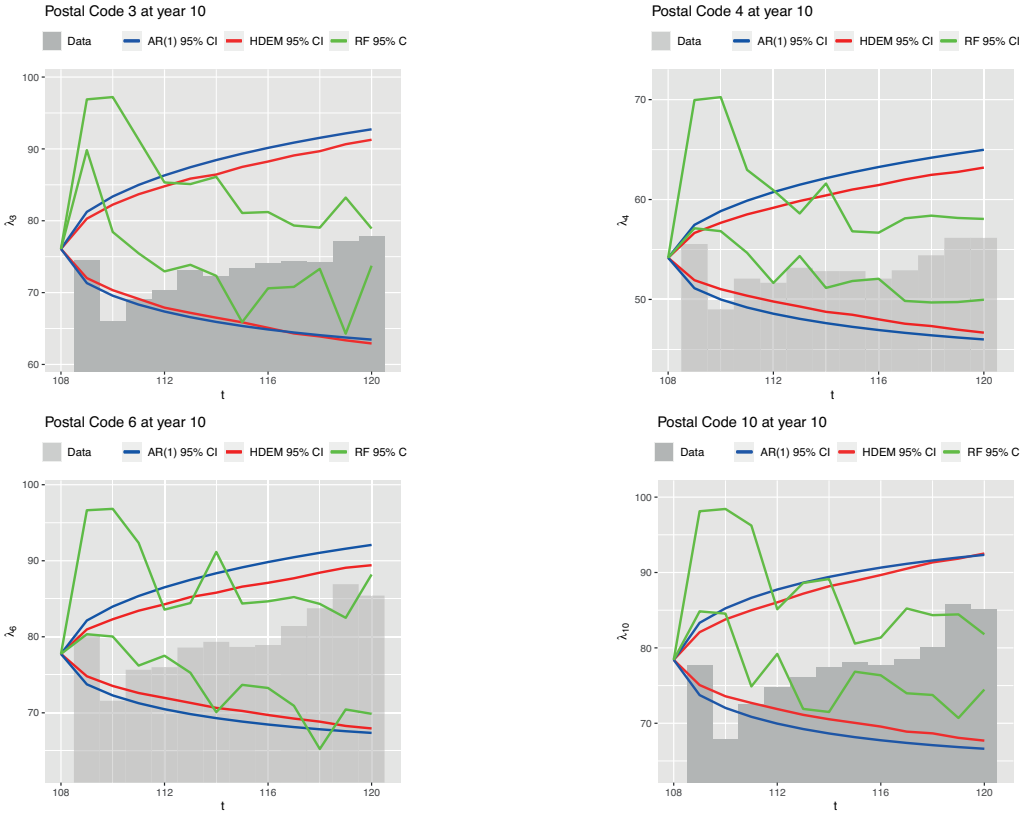

Figure 9. *Comparison between the credibility intervals of the proposed model (HDEM), the AR(1), and the Random Forest (RF) for one-year predictions for year 10 and some selected postal codes (3, 4, 6, 10).*

It is important to stress that one of the key strengths of our model is its ability to account for the spatial correlation between different areas. This feature is crucial for

capturing the interactions between neighboring regions, which are often an essential component in the spatial dynamics of crime. Although an AR structure may show strong predictive capabilities in certain contexts, it lacks the spatial interpretability and granularity that our model offers (also true in the case of an RF). This makes our approach particularly valuable for criminologists and law enforcement agencies who want not only predict but also gain a deeper understanding of the spatial patterns and underlying dynamics of crime.

We additionally calculated the root mean squared error (RMSE) of the predictions for each model by comparing the estimates with real values for the last year. By doing this across the 10 zones, we obtain an error for each of them. In Table 1, we show the average error over these 10 zones, as well as the 90% confidence interval of these errors. The results highlight that the RMSE of our model is lower than that of the other two classical models, supporting the better precision and reliability of the HDEM approach over traditional models.

Model	Average error over 10 zip codes	90%
HDEM	15.16359	[7.63, 26.61]
AR(1)	15.38856	[7.97, 26.80]
RF	27.06137	[17.71, 39.59]

Table 1. *Root mean square error (RMSE) for each model based on predictions during the final* 10*th year. The training period spans the previous years (*< 10*).*

4. Discussion and conclusions

We have detailed a methodology to deal with crime data, based on spatio-temporal stochastic differential equations. The data were point-referenced or distributed on patches. From a stochastic partial differential equation for the intensity with temporal log-Gaussian evolution, we obtained stochastic differential equations with correlated Wiener noises. The correlations were stochastic processes too, with a transformed log-Gaussian distribution. We have proposed a hierarchical Bayesian structure for joint inference, and model fitting was conducted for the lattice data in Valencia, which contains a sufficiently large number of events. Time series on the irregular grid were fitted for the past (interpolation) and for the future (prediction, either for a long time window or on an instant-by-instant basis). Our work combined ideas from stochastic differential equations, spatial statistics, time series, and quantitative finance. It is the first attempt to incorporate the use of stochastic partial differential equations in mathematical/statistical criminology with real data, as a direct modeling tool. Other areas of application could be epidemiology or ecology.

Our proposal is very much focused on modeling the dynamics of the stochastic intensity function through differential equations rather than through an underlying Gaussian process. In this way, we can better control the stochastic term through the drift param-

eter $\mu()$ that captures the growth rate, and the volatility parameter $\sigma()$ that captures the magnitude of fluctuations and the uncertainty of future values. These two parameters are inspired by the evolution of crime incidence and social crime, where there is a rate for the growth of criminality and a volatility for random fluctuations. Note that our approach goes more in the line of Zammit-Mangion et al. (2012), which can be further embedded into an LGCP framework in the context of stochastic integro-difference equations. This is not the usual framework in LGCPs. Classical and widely used LGCPs are based on an underlying Gaussian process (GP) and the second-order properties are inherited from the covariance structure of this GP. Using covariances in space and time is much less flexible, carries a computational burden, and does not model the dynamics of the process as a differential equation does. A final aspect is that with our approach we can model stochastic correlations varying in time, something that becomes very troublesome in classical LGCPs.

The stochastic log-normal model has been of use in finance for stock-price evolution with correlations. The time series of crime studied in this paper have similar fluctuations to those encountered in financial modeling, hence our model building. An important feature of our proposal is that parameter estimation is conducted within the Bayesian paradigm, enabling the delineation of the uncertainty in each involved parameter. We are not aware of similar differential-equation models in criminology, with such a statistical treatment.

Further investigation of this type of models would be of interest when the number of sampling sites or regions of study augments, producing a high increase in the number of parameters. Convergence and computational times are issues to be solved to make these models useful in practice. The case of point patterns, for which locations and times arise randomly, is left for future research. The analysis of point patterns would help to identify crime hotspots within the zip codes. Otherwise, we are constrained to uniformly distributed events within the spatial districts, with the quantity given by the Poisson realization of the aggregate intensity. Finally, there is still work to be done for the inclusion of covariates, probably through link functions on the drift coefficients, to seek higher predictability (with no overfitting): more accurate mean values from a pointwise sense and narrower (less uncertain) probabilistic intervals. Overfitting occurs when models are excessively complex and match the data too perfectly (Duan et al., 2009, Footnote 3), with some parameters that may be unidentifiable (Smith, 2013). Note that in Figure 8 the model does an effective and practical job when predicting month by month for real data, and the incorporation of covariates shall be of use to expand the prediction window.

Funding

This paper has been supported by grants UJI-B2021-37 from Universitat Jaume I and PID2022-141555OB-I00, PID2023-146836NB-I00 from Spanish Ministry of Science.

Data availability statement

The data analyzed in this study are available from the authors upon reasonable request.

Disclosure statement

The authors declare that there is no conflict of interests regarding the publication of this article.

References

Abbas, S., Tripathi, J. P. and Neha, A. A. (2017). Dynamical analysis of a model of social behavior: criminal vs non-criminal population. Chaos Solitons & Fractals, 98, 121–129.

Acedo, L., Díez-Domingo, J., Morano, J. A. and Villanueva R. J. (2010). Mathematical modelling of respiratory syncytial virus (RSV): vaccination strategies and budget applications. Epidemiology & Infection, 138(6), 853–860.

Allen, E. (2007). Modeling With Itô Stochastic Differential Equations. Springer Science & Business Media, Dordrecht, Netherlands.

Banerjee, S., Carlin, B. P. and Gelfand A. E. (2014). Hierarchical Modeling and Analysis for Spatial Data, 2nd Ed. Chapman & Hall/CRC, Boca Raton.

Brauer, F. (2008). Compartmental models in epidemiology. In: Brauer, F., van den Driessche, P. and Wu, J. (eds), Mathematical Epidemiology. Springer, Berlin, Heidelberg.

Braumann, C. A. (2007). Itô versus Stratonovich calculus in random population growth. Mathematical Biosciences, 206(1), 81–107.

Brooks, S., Gelman, A., Jones G. L. and Meng X. L. (2011). Handbook of Markov Chain Monte Carlo, 1st Ed. Chapman & Hall/CRC, Boca Raton.

Burgess, R. L. and Akers, R. L. (1966). A differential association-reinforcement theory of criminal behavior. Social Problems, 14(2), 128–147.

Calatayud, J., Jornet, M. and Mateu, J. (2025a). Spatial modeling of crime dynamics: Patch and reaction-diffusion compartmental systems. Mathematical Methods in the Applied Sciences, 48(7), 7440-7459.

Calatayud, J., Jornet, M. and Mateu, J. (2023a). Modeling noisy time-series data of crime with stochastic differential equations. Stochastic Environmental Research and Risk Assessment, 37, 1053–1066.

Calatayud, J., Jornet, M. and Mateu, J. (2023b). Spatio-temporal stochastic differential equations for crime incidence modeling. Stochastic Environmental Research and Risk Assessment, 37, 1839–1854.

Calatayud, J., Jornet, M. and Mateu, J. (2025b). A dynamical mathematical model for crime evolution based on a compartmental system with interactions. International Journal of Computer Mathematics, 102(1), 44-59.

Cao, Y., Dong, K., Siercke, B. and Wilber, M. (2013). Final Report: Crime Modeling. UCLA, LA, USA. https://www.math.ucla.edu/~bertozzi/WORKFORCE/REU%202013/Crime%20Fighters%20%5BGroup%20Awesome%5D/FinalReport.pdf. Last accessed 14 July 2023.

Cervelló, R., Cortés, J. C., Santonja, F. J. and Villanueva R. J. (2014). The dynamics over the next few years of the Spanish mobile telecommunications market share: a mathematical modelling approach. Mathematical and Computer Modelling of Dynamical Systems, 20(6), 557–565.

Davison, A.C. and Hinkley, D.V. (1997). Bootstrap Methods and Their Application. Cambridge University Press, London.

Dorini, F. A. and Sampaio, R. (2012). Some results on the random wear coefficient of the Archard model. Journal of Applied Mechanics, 79(5), 051008–051014.

Duan, J. A., Gelfand, A. E. and Sirmans, C.F. (2009). Modeling space-time data using stochastic differential equations. Bayesian Analysis, 4(4), 733–758.

Esiri, M. O. (2016). The influence of peer pressure on criminal behaviour. Journal of Humanities and Social Science, 21(1), 08–14.

Evans, L. C. (2012). An Introduction to Stochastic Differential Equations. American Mathematical Society, New York.

Gelfand, A. E., Diggle, P. J., Fuentes, M. and Guttorp, P. (2010). The Handbook of Spatial Statistics. Chapman Hall, Boca Raton.

González-Parra, G., Chen-Charpentier, B. and Kojouharov, H. V. (2018). Mathematical modeling of crime as a social epidemic. Journal of Interdisciplinary Mathematics, 21(3), 623–643.

Harkins, S. G., Williams, K. D. and Burger, J. (2017). The Oxford Handbook of Social Influence. Oxford University Press, UK.

Jane White, K. A., Campillo-Funollet, E., Nyabadza, F., Cusseddu, D., Kasumo, C., Imbusi, N. M., Ogesa Juma, V., Meir, A. J., Marijani, T. (2021). Towards understanding crime dynamics in a heterogeneous environment: A mathematical approach. Journal of Interdisciplinary Mathematics, 24(8), 2139–2159.

Kolokolnikov, T., Lloyd, D. J. B. and Short M. (2019). Mathematical Criminology and Security. Workshop, Banff International Research Station, Banff, Canada, 17th March 2019 – 22nd March 2019, pp. 1–7.

Kolokolnikov, T., Ward, M. J. and Wei, J. (2014). The stability of steady-state hot-spot patterns for a reaction-diffusion model of urban crime. Discrete and Continuous Dynamical Systems - B, 19(5), 1373–1410.

Koss, L. (2019). SIR models: differential equations that support the common good. CODEE Journal, 12(1), article 6.

Lacey, A. A. and Tsardakas, M. N. (2016). A mathematical model of serious and minor criminal activity. European Journal of Applied Mathematics, 27(3), 403–421.

Lamberton, D. and Lapeyre, B. (2011). Introduction to Stochastic Calculus Applied to Finance, 2nd Ed. Chapman & Hall / CRC press, London.

Lesaffre, E. and Lawson, A. B. (2012). Bayesian Biostatistics. Wiley, New York.

Lloyd, D. J. and O'Farrell, H. (2013). On localised hotspots of an urban crime model. Physica D: Nonlinear Phenomena, 253, 23–39.

Lord, G. J., Powell, C. E. and Shardlow, T. (2014). An Introduction to Computational Stochastic PDEs. Cambridge Texts in Applied Mathematics, Cambridge University Press, New York.

Mao, X. (2007). Stochastic Differential Equations and Applications. Elsevier, Cambridge.

McMillon, D., Simon, C. P. and Morenoff, J. (2014). Modeling the underlying dynamics of the spread of crime. PLoS ONE, 9(4), e88923.

Murray, J. D. (2002). Mathematical Biology I. 3rd Ed. Springer, New York.

Neal, R. M. (2003). Slice sampling. Annals of Statistics, 31(3), 705–767.

R Core Team. (2023). R: A language and environment for statistical computing, Version 4.3.2. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rackauckas, C. (2014). An Intuitive Introduction For Understanding and Solving Stochastic Differential Equations. Course notes.

Rodriguez, N. and Winkler, M. (2022). On the global existence and qualitative behaviour of one-dimensional solutions to a model for urban crime. European Journal of Applied Mathematics, 33(5), 919–959.

Santonja, F. J., Sánchez, E., Rubio, M. and Morera, J. L. (2010). Alcohol consumption in Spain and its economic cost: a mathematical modeling approach. Mathematical and Computer Modelling, 52(7–8), 999–1003.

Schiesser, W. E. (2019). Spatiotemporal Modeling of Influenza. Partial Differential Equation Analysis in R. Synthesis Lectures on Biomedical Engineering. Morgan & Claypool Publishers, USA.

Short, M. B., d'Orsogna, M. R., Pasour, V. B., Tita, G. E., Brantingham, P. J., Bertozzi, A. L. and Chayes, L. B. (2008). A statistical model of criminal behavior. Mathematical Models and Methods in Applied Sciences, 18(01), 1249–1267.

Short, M., Bertozzi, A. and Brantingham, P. (2010a). Nonlinear patterns in urban crime: Hotspots, bifurcations, and suppression. SIAM Journal on Applied Dynamical Systems, 9(2), 462–483.

Short, M. B., Brantingham, P. J., Bertozzi, A. L., Tita, G. E. (2010b). Dissipation and displacement of hotspots in reaction-diffusion models of crime. Proceedings of the National Academy of Sciences, 107(9), 3961–3965.

Smith, R.C. (2013) Uncertainty Quantification: Theory, Implementation, and Applications. SIAM, Philadelphia.

Song, B., Castillo-Garsow, M., Ros-Soto, K. R., Mejran, M., Henso, L. and Castillo-Chávez, C. (2006). Raves, clubs and ecstasy: the impact of peer pressure. Mathematical Biosciences and Engineering, 3(1), 249–266.

Sooknanan, J. and Comissiong, D.M. (2017). When behaviour turns contagious: the use of deterministic epidemiological models in modeling social contagion phenomena. International Journal of Dynamics and Control, 5(4), 1046–1050.

Spanish Official State Bulletin (BOE), number 67, decree 463/2020, reference BOE-A-2020-3692, 14th March 2020. https://www.boe.es/eli/es/rd/2020/03/14/463/con

Srivastav, A. K., Athithan, S. and Ghosh, M. (2020). Modeling and analysis of crime prediction and prevention. Social Network Analysis and Mining, 10(1), 1–21.

Teng, L., Ehrhardt, M. and Günther, M. (2016). Modelling stochastic correlation. Journal of Mathematics in Industry, 6, 1–18.

Tse, W.H. and Ward, M.J. (2016). Hotspot formation and dynamics for a continuum model of urban crime. European Journal of Applied Mathematics, 27(3), 583–624.

van den Driessche, P. (2008). Compartmental models in epidemiology. In: Brauer, F., van den Driessche, P. and Wu, J., (eds), Mathematical Epidemiology. Springer, Berlin, Heidelberg.

van Emmerich, C. (2006). Modelling correlation as a stochastic process. Preprint 06/03, University of Wuppertal.

Voit, J. (2010). The Statistical Mechanics of Financial Markets, 3rd Ed. Springer, Berlin, Heidelberg.

Wang, Q., Wang, D. and Feng, Y. (2020). Global well-posedness and uniform boundedness of urban crime models: One-dimensional case. Journal Of Differential Equations, 269(7), 6216–6235.

White, E. and Comiskey, C. (2007). Heroin epidemics, treatment and ODE modeling. Mathematical Biosciences, 208(1), 312–324.

Wolfram Research, Inc. (2020). Mathematica, Version 12.1. Champaign, IL.

Wu, J. (2008). Spatial structure: Partial differential equations models. In: Brauer, F., van den Driessche, P. and Wu, J. (eds), Mathematical Epidemiology. Springer, Berlin, Heidelberg.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y., Yuan, M. L., Zhang, Y. L., Dai, F. H., Liu, Y., Wang, Q. M., Zheng, J. J., Xu, L., Holmes, E. C. and Zhang, Y. L. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 579(265), 265–269.

Xiu, D. (2010). Numerical Methods for Stochastic Computations: A Spectral Method Approach. Cambridge Texts in Applied Mathematics. Princeton University Press, New York.

Zammit-Mangion, A., Dewar, M., Kadirkamanathan, V. and Sanguinetti, G. (2012). Point process modelling of the Afghan War Diary. Proceedings of the National Academy of Sciences, 109(31), 12414–12419.

Zammit-Mangion, A. and Wikle, C.K. (2020). Deep integro-difference equation models for spatio-temporal forecasting. Spatial Statistics, 37, 100408.

SORT 49 (2) July-December 2025, 179-212 DOI: 10.57645/20.8080.02.27

Optimism correction of the area under the ROC curve, with missing data

Susana Rafaela Martins[1], María del Carmen Iglesias-Pérez[2] and Jacobo de Uña-Alvarez[2]

Abstract

The area under the ROC curve (AUC) plays an important role in the study of the predictive capacity of regression models. It is well known that an inflated AUC may result when the same data are used for training and testing the model. In this paper optimism correction of the AUC in the presence of missing data is investigated. Complete case analysis, inverse probability weighting and multiple imputation are employed to address the issue of missing data. For each of these approaches, split-sample, K-fold cross-validation and leave-one-out cross-validation are employed to correct for the optimism of the AUC. The methods are compared through intensive Monte Carlo simulations in the particular setting of binary regression. Results suggest that all estimators are consistent with the exception of complete case analysis, which may be biased when missing is not completely at random. In general, a combined application of multiple imputation and leave-one-out cross-validation is recommended.

MSC: *62G05, 62G09, and 92B15.*

Keywords: *Cross-validation, logistic regression, missing values, multiple imputation, prediction.*

1. Introduction

Predictive models are widely used in different areas. They are used to evaluate credit risk, evaluate fraud hypotheses, analyse the performance of a machine or identify the individual who is at greatest risk of illness (Fawcett, 2006; Pepe, 2003; Wishart et al., 2012; Quintana et al., 2014; Garcia-Gutierrez et al., 2017). A critical step in the application of predictive models is the assessment of their predictive ability, that is, the performance when applied to new individuals. In this work, we focus on binary regression models with missing values and on the study of their discrimination capacity (Steyerberg et al., 2001, 2003).

[1] Escola Superior de Desporto e Lazer, Instituto Politécnico de Viana do Castelo, Portugal.

[2] Department of Statistics and OR, Universidade de Vigo, Spain.

Received: February 2025

Accepted: July 2025

The Area Under the ROC Curve (AUC) is a discrimination measure that evaluates the overall ability of a model to correctly classify observations into two distinct classes. It represents an accuracy index for a classifier. When its value is 1, it indicates perfect discrimination capacity. If this value is 0.5, the discrimination capacity is null (Fan, Upadhye and Worster, 2006; Pepe, 2003).

In general, when the same data are used for training and testing the model the AUC is overestimated. In other words the AUC estimate can be optimistic; this is so-called apparent AUC (APP AUC). Several methods to correct for this optimism have been proposed in the context of complete data. Such methods include Split-sample (SS), K-fold cross-validation (KF), Leave-one-out (LOO) cross-validation or Bootstrap (Iparragirre, Barrio and Rodriguez-Alvarez, 2019; Airola et al., 2011; Smith et al., 2014; Austin and Steyerberg, 2017). However, in practice, the data could have missing information. Missing data in the response variable or in the covariables may have an impact in the performance of prediction models. In particular, when the individuals with missing information are removed from the sample, the so-called Complete Case (CC) analysis, the estimators of the regression coefficients and of discrimination measures such as the AUC may be biased. There exist some estimators of the AUC proposed to adjust biases caused by missing values. Inverse Probability Weighting (IPW) is commonly used to correct the selection bias when the analysis is restricted to cases with complete information. Alternatively, the analysis could be improved by using other methods for missing data, such as Multiple Imputation (MI) (Li et al., 2021; Cho, Matthews and Harel, 2019).

To sum up, corrections for the optimism of the AUC together with methods to properly handle missing data are needed in practice. As far as we know, only few studies in the literature addressed these two problems. Wahl et al. (2016) focus on MI combined with Bootstrap and KF methods; however, they do not include LOO method, nor consider other methodologies for missing data such as IPW or CC. On the other hand, Mertens, Banzato and de Wreede (2020) investigate the problem of calibrating a prediction model in the presence of missing data using MI. Nevertheless, the problem of the estimation of the AUC is not considered in this latter paper.

The goal of this work is the correction of optimism in the estimation of AUC in the presence of missing values. We compare various methodologies for handling missing data - MI, CC, and IPW – combined with different optimism correction methods – SS, KF, and LOO. In this sense, this study provides novel contributions to the topic of correcting the optimism of the AUC with missing data. Specifically, the performance of IPW with optimism correction is investigated for the first time. Also, the benefits of LOO when correcting for the optimism of the AUC in the missing data setting are explored.

The rest of the paper is organized as follows. In Chapter 2 the usual estimators for the AUC, corrections for their optimism with complete data, and existing adaptations of the empirical AUC to the context of missing data are presented. In Chapter 3 the methodologies to correct the optimism of AUC are adapted to missing data. The methods are compared through intensive Monte Carlo simulations in the particular setting of logistic

regression in Chapter 4. In Chapter 5 we present two real data illustrations. Finally, the main conclusions or our study are reported in Chapter 6.

2. Methods for complete data

2.1. AUC for Binary Regression

AUC is the most used metric to evaluated the performance of classification models, representing the area under the ROC curve. The ROC curve is a graphical representation of the trade-off between the true positive rate and the false positive rate of a binary classifier as its discrimination threshold is varied. In the general binary regression model we denote by Y the response variable, X the corresponding q-dimensional vector of covariates, and $\{(X_i,Y_i), 1 \leq i \leq n\}$ a random sample of (X,Y).

For a given classifier $p(X)$, AUC is the probability that $p(X)$ takes a larger value for an individual randomly drawn from the diseased population compared to an individual sampled from the healthy population:

$$AUC = P(p(X_1) < p(X_2)|Y_1 = 0, Y_2 = 1),$$

where $Y_1 = 0$ and $Y_2 = 1$ represent the nondiseased and diseased population, respectively.

Alternatively, AUC can be expressed as the expected value of an indicator function:

$$AUC = E(I(p(X_1) < p(X_2))|Y_1 = 0, Y_2 = 1)$$

where the indicator function $I(.)$ takes the value 1 if its argument is true or 0 otherwise (Pepe, 2003).

Without loss of generality, consider the logistic regression model $Y_i \sim Bernoulli(p(X_i))$ where

$$p(X_i) = P(Y_i = 1|X_i) = \frac{\exp(\beta^{\intercal} X_i)}{1+\exp(\beta^{\intercal} X_i)} \tag{1}$$

and β is the unknown vector of regression coefficients, which can be estimated from the data leading to the feasible predictor $\hat{p}(X)$. Then, the AUC estimator is given by

$$\widehat{AUC} = \frac{1}{n_0 n_1} \sum_{i \in D_0} \sum_{j \in D_1} [I(\hat{p}(X_i) < \hat{p}(X_j)) + 0.5 I(\hat{p}(X_i) = \hat{p}(X_j))] \tag{2}$$

where $D_0 = \{i|Y_i = 0\}$ and $D_1 = \{i|Y_i = 1\}$ represent the nondiseased and diseased individuals in the sample, respectively, and where n_0 and n_1 are their corresponding cardinalities. The empirical AUC in (2) is indeed equal to the two-sample Mann-Whitney-Wilcoxon statistic.

2.2. Correction for the optimism of the AUC

The empirical AUC could be inflated if the same data set is used to fit and to test the model. The most commonly used methods for correcting the optimism of the AUC with

complete data are briefly reviewed below. See, for instance, Iparragirre et al. (2019) and references therein for further details.

2.2.1. Split-sample (SS) cross-validation

In SS the sample is randomly divided into two subsamples: *training (train)* sample and *test* sample. Subsequently, the regression coefficients are estimated from the training sample. Using these coefficients, the prediction of the response variable $\hat{p}(X_i)$ for the *test* sample is calculated according to (1). Taking into account these estimated probabilities and the Y_i values in the *test* sample, the AUC is calculated according to equation (2). SS cross-validation is frequently used with two samples with equal size.

2.2.2. K-fold (KF) cross-validation

In KF cross-validation, the sample is divided into K subsamples of approximately similar sizes. The sub-sample S_k, $1 \leq k \leq K$, is the $test_k$ sample and the set with all of the others sub-samples is the $train_k$ sample. The regression model is estimated with the $train_k$ sample and then it is used to compute $\hat{p}(X_i)$ for the X_i in the $test_k$ sample. So, the respective AUC is calculated. This procedure is repeated for all S_k, resulting in K AUCs. The corrected AUC is calculated using the mean of the K-AUCs. In general, $K = 10$ is the most commonly used in the literature.

2.2.3. Leave-one-out (LOO) cross-validation

In LOO one observation is omitted from the initial set and the regression model is fitted from the remaining observations. The estimated model is used to compute $\hat{p}(X_i)$ for the X_i that was left out. This procedure is repeated for all observations in the sample, and the AUC is calculated comparing the estimated probabilities to the corresponding Y_i.

2.3. AUC with missing values

In the missing data setting the empirical AUC in (2) is no longer available and, therefore, some adjustments are needed. Given a vector (X_i, Y_i) of covariables X_i and variable response Y_i, in line with the approaches of Molenberghs and Kenward (2007) and Chen, Wan and Zhou (2015), we consider the corresponding vector (Z_i, Z_i^{mis}), where Z_i is a d-dimensional vector with $0 < d \leq q$ that is observed for all i's, while Z_i^{mis} represents the variables that may or may not be available for some i's. Note that the response variable will be included in the vector Z_i^{mis} when its value is missing for some individuals. Let R_i be the indicator of missing values, meaning,

$$R_i = \begin{cases} 1, & if \ \ Z_i^{mis} \ \ is \ \ observed \\ 0, & if \ \ Z_i^{mis} \ \ is \ \ missing \end{cases} \tag{3}$$

The data are considered missing completely at random (MCAR) when the probability of missing values is independent of both Z_i and Z_i^{mis}. Consequently, under

MCAR assumption, the probability of missingness is given by $P(R_i = 0|X_i, Y_i) = P(R_i = 0|Z_i, Z_i^{mis}) = P(R_i = 0)$. On the other hand, missing at random (MAR) occurs when the probability of missing values depends on Z_i. Under MAR assumption, the probability of missingness is $P(R_i = 0|X_i, Y_i) = P(R_i = 0|Z_i)$. Another mechanism of missing values is missing not at random (MNAR). Under MNAR the probability of missingness depends on both Z_i and Z_i^{mis}. That is, the probability of missingness is $P(R_i = 0|X_i, Y_i) = P(R_i = 0|Z_i, Z_i^{mis})$. MNAR scenarios are difficult since the missing mechanism depends on variables which are not always available, so external information may be needed in order to proceed.

2.3.1. Complete Case (CC)

CC analysis simply proceeds by deleting from the sample the individuals that have missing information. Following the same idea presented in Li et al. (2021), the expression of the CC version of the AUC is

$$\widehat{AUC}_{cc} = \frac{\sum_{i\in D_0}\sum_{j\in D_1}[I(\hat{p}(X_i) < \hat{p}(X_j)) + 0.5I(\hat{p}(X_i) = \hat{p}(X_j))]R_iR_j}{\sum_{i\in D_0}R_i\sum_{j\in D_1}R_j} \tag{4}$$

The sums $\sum_{i\in D_0}R_i$ and $\sum_{j\in D_1}R_j$ are the number of observations without missing values that are nondiseased and diseased, respectively. The estimator (4) is consistent under MCAR. However, it may be inconsistent in MAR scenarios; see for instance Li et al. (2021).

2.3.2. Inverse Probability Weighting (IPW)

IPW uses the inverse of the estimated probability that an individual has complete information to weight each observation and thus to correct the potential selection bias. Let $W_i = 1/P(R_i = 1|X_i, Y_i)$ be the inverse probability of the observation to be complete. Aligned with the idea of Li et al. (2021), we use logistic regression to build a model for $P(R_i = 1|X_i, Y_i) = P(R_i = 1|Z_i)$ conditional on the fully observed variates (under MAR assumption), and then to obtain the weight estimates $\hat{W}_i$. Then, the AUC IPW estimator is

$$\widehat{AUC}_{ipw} = \frac{\sum_{i\in D_0}\sum_{j\in D_1}[I(\hat{p}(X_i) < \hat{p}(X_j)) + 0.5I(\hat{p}(X_i) = \hat{p}(X_j))]R_i\hat{W}_iR_j\hat{W}_j}{\sum_{i\in D_0}R_i\hat{W}_i\sum_{j\in D_1}R_j\hat{W}_j} \tag{5}$$

The sums $\sum_{i\in D_0}R_iW_i$ and $\sum_{j\in D_1}R_jW_j$ are the weighted observations without missing values that are nondiseased and diseased, respectively. When the MAR-logistic model for the weights W_i is correctly specified, the estimator (5) is consistent (See Section 2.5 in Li et al. (2021)).

2.3.3. Multiple Imputation (MI)

MI replaces missing data with imputed values, resulting in multiple 'completed' datasets. Several approaches exist for performing imputation, such as the multivariate normal model developed by Schafer (1997) or the full conditional specification (FCS) method proposed by Raghunathan et al. (2001); van Buuren (2007), also known as "chained equations". The FCS is based on distributions of fully observed variables and is one of the most used in multiple imputation to estimate the distribution of partially observed variables (Carpenter and Smuk, 2021). FCS is expected to be consistent when the involved chained equations are correctly specified. The good practical behaviour of MI has been widely studied in the literature; see for instance Zhu and Raghunathan (2015); Carpenter and Smuk (2021).

In the case of AUC, M imputations are performed and M datasets with no missing data are obtained. For each of these sets, the respective AUC is estimated, AUC_m, according to the equation (2). The AUC resulting from this methodology, $\widehat{AUC}_{mi}$, is estimated by the average of the M AUCs:

$$\widehat{AUC}_{mi} = \frac{1}{M}\sum_{m=1}^{M} AUC_m. \tag{6}$$

In this paper, we adopted the fully conditional specification approach and selected $M = 5$ imputations, following common practice and the recommendation of van Buuren (2018), who notes that increasing M beyond 5 is unlikely to alter the substantive conclusions. Since MI yields several completed datasets, we report the mean AUC across all imputed sets. Note that the objective in this study is to evaluate and compare the predictive performance of different methods for handling missing data and optimism correction, specifically, using the AUC as the parameter of interest. To estimate the AUC under MI, we adopt the common approach of computing it separately for each imputed dataset and then averaging the resulting AUCs. This strategy is consistent with recommendations in the literature (e.g., Wahl et al. (2016); Mertens et al. (2020)), and is preferred when the interest lies in assessing model performance rather than interpreting coefficients. It is worth noting that if one were interested in reporting a final model for implementation, the appropriate approach would be to pool the regression coefficients across imputations using Rubin's rules (Rubin, 1987). A single model could then be derived for interpretation purposes, and its AUC could be computed. However, this is not the focus of the current study.

3. Correction for the optimism of the AUC with missing values

Two general approaches to correct for the optimism of the AUC with missing data are possible. The first one corrects first for optimism and then deals with the missingness issue. This approach is feasible when using SS cross-validation or KF cross-validation, among other methods. However, the approach fails for LOO cross-validation, since prediction for an individual observation is not possible when some covariates are missing.

The second approach solves first the missingness issue and then proceeds to correct for optimism. This approach works for all the methods, and it will be employed in our research.

3.1. SS cross-validation for CC, IPW and MI methods

3.1.1. SS cross-validation for CC analysis

Taking into account the CC methodology, the set of complete observations is considered and this set is partitioned in half into *trainc* and *testc* subsamples. Using the set *trainc* the regression parameters are estimated and, from this estimated model, predictions are obtained for the *testc* set. Using the response values of *testc*, and their predictions, the AUC, which we call AUC_{cc-ss}, is calculated using (4).

3.1.2. SS cross-validation for IPW

In the IPW methodology, one estimates the weights, $W_i = 1/P(R_i = 1|Z_i)$, by plugging in a consistent estimator for the non-missing probability $P(R_i = 1|Z_i)$. Then, the dataset is divided into two sets with the same size, *trainc* and *testc*, and the respective pre-estimated weights are considered.With the *trainc* data set and the respective pre-calculated weights, a weighted logistic model is built and predictions for the *testc* sample are obtained. The corresponding AUC is calculated taking into account formula (5) and using the response values, pre-estimated weights and predicted values of the *testc* set. To be more specific, $X_i, X_j, W_i, W_j, R_i, R_j, D_0$ and D_1 are related to *testc*, while $\hat{p}$ is estimated using weighted logistic regression with *trainc* sample, and evaluated in the *testc* sample.

3.1.3. SS cross-validation for MI

In the case of MI, M (we take $M = 5$ in the simulations below) imputations are performed to construct M complete datasets. For each imputation, one splits the full set into two, *mictrain* and *mictest* subsets. Then, the regression coefficients are estimated from each mictrain set. Using the estimated regression model, the predicted values of the respective *mictest* sets are calculated and the corresponding AUCs are obtained from (2). The final AUC, AUC_{mi}, is defined as the mean of the M AUCs obtained.

3.2. KF cross-validation for CC, IPW and MI methods

3.2.1. KF cross-validation for CC analysis

Taking into account the CC approach, only the complete observations are considered. The set of complete observations is divided into K $(K = 10)$ sets and the sets $trainc_k$ and $testc_k$ are obtained as described in Section 2.2.2. Using the set $trainc_k$ the regression parameters are estimated. Based on this estimated model, predictions are obtained for

the $testc_k$ sample. With the true outcomes of each $testc_k$ and their predictions, each AUC is calculated, AUC_{cc_k}, according to (4). Finally the K, AUC_{cc_k} are averaged.

3.2.2. KF cross-validation for IPW

The data set is divided into K (we take $K = 10$ in the simulations below) sets, $train_k$ and $test_k$. The method proceeds first as described in Section 3.1.2 for each of the K folds of the dataset and then the final AUC is obtained by averaging.

3.2.3. KF cross-validation for MI

The method proceeds first as described in Section 3.1.3 for each of the K folds of the dataset and then the final AUC is obtained by averaging.

3.3. LOO cross-validation for CC, IPW and MI methods

3.3.1. LOO cross-validation for CC analysis

LOO cross-validation for CC proceeds just as described in 2.2.3 considering only the complete observations.

3.3.2. LOO cross-validation for IPW

In the IPW method, the weights are first estimated for all individuals in the sample. These weights correspond to the inverse of the estimated probability of having complete data, typically obtained by fitting a logistic model to the missingness indicator R_i, conditional on the fully observed variables (i.e., estimating $\hat{W}_i = 1/\hat{P}(R_i = 1|Z_i)$). After estimating the weights, leave-one-out cross-validation is applied. Each observation i, $1 \leq i \leq n$, is considered as the *test* set, while the remaining $n-1$ observations form the corresponding train set. Using the train set and the respective pre-estimated weights, a weighted logistic regression model is fitted to predict the outcome variable Y. The fitted model is then used to compute the predicted probability for the omitted observation i. This process is repeated for all n observations, resulting in a set of n predicted probabilities. Finally, the AUC is computed using these predictions and the observed outcomes, according to equation (5).

3.3.3. LOO cross-validation for MI

For the MI methodology, missing data are imputed and M complete samples are constructed. For each of this samples train and test sets are defined. Each observation i, $1 \leq i \leq n$ is considered the test sample, and the original set without this observation i is the train. The regression model is estimated from each of the train samples. Each estimated regession model is used to obtain the prediction of the outcome for the corresponding test sample. This results in M AUCs, one for each imputed dataset. The final AUC is obtained by averaging.

4. Simulation study

In this section the performance of optimism correction methods with missing values, as introduced in Section 3, is investigated through simulations. The setting is that of logistic regression. The goal is to identify the best methods to correct for missing values and for the optimism of the AUC. We compare the AUC estimated by each combination of methods to the "true" out-of-sample AUC associated to the fitted logistic regression models. As mentioned, the combination of methods involves a method to correct for data missingness (*mm*) and a method to correct for the optimism of the AUC (*om*). The "true" out-of-sample AUC represents the true discriminatory ability of the models, according to a particular missing method, when applied to new data without missingness. We consider various factors that might affect the methods' performance, including the sample size and disease prevalence as in Iparragirre et al. (2019). Inspired by Li et al. (2021), we considered different scenarios of missingness. Details are given in the next section.

4.1. Simulation design

In the simulation study two independent samples $\{X_i, Y_i\}_{i=1}^{n}$ and $\{X_l, Y_l\}_{l=1}^{N}$, *ndata* and *bigdata* say, from the population vector (X, Y) were generated, where Y was the binary response variable and X was a vector of eight covariates. The steps to simulate each (X_i, Y_i) were the following.

- Draw a Bernoulli (*prev*) variable, η_i,
- Given η_i, draw X_i from a multivariante Normal distribution with independent components with standard deviation 0.6 as follows:
  - If $\eta_i = 0$ the components of X_i were zero mean.
  - If $\eta_i = 1$ the vector mean of X_i was $(0.6, 0.55, 0.5, 0.45, 0.4, 0.3, 0.25, 0.2)$.
- Draw Y_i from a Bernoulli(π_i) distribution, where $\pi_i = \frac{\exp(\beta^T X_i)}{1+\exp(\beta^T X_i)}$, where β is the true vector of regression coefficients.

The prevalence values (*prev*) in the simulations were 0.1, 0.2 and 0.5. The true regression coefficient vector was
$\beta = [-2.5082, 0.5625, 0.4375, 0.3125, 0.1875, 0.0625, -0.1875, -0.3125, -0.4375]$.
Three different sizes were considered for *ndata*, $n = \{250, 500, 2000\}$. The size of *bigdata* was $N = 50000$ and $T = 500$ Monte Carlo trials were performed. Our simulation design resembles that in Iparragirre et al. (2019). A novelty here is the data missingness, which was introduced for a single covariate, namely X_1, according to different scenarios:

- S1: MCAR with missing probability $P(R = 0) = 0.5$.

- S2: MAR with missing probability depending on X_2 and X_3: $P(R=0|X_2,X_3) = 1/(1+exp(-0.5+2X_2-X_3)$
- S3: MAR with missing probability depending on X_2 and Y: $P(R=0|X_2,Y) = 1/(1+exp(-0.5+2X_2+1.5Y))$

On average, the probability of missingness in the MAR scenarios was approximately 0.5. To fit the logistic regression model we used the `glm` function of `R` with binomial family (R Core Team, 2013). The weights for IPW were also computed by fitting a logistic regression model with R function `glm`. For MI we used the function `mice` of the package in `R` with same name (van Buuren and Groothuis-Oudshoorn, 2011). We used $M=5$ imputed datasets, applying the `norm` method for imputing variable X_1. To estimate the AUC, equations (4), (5) and (6) were implemented.

We evaluated the AUC on each Monte Carlo trial and then we computed the Monte Carlo average, bias and mean square error (MSE) as follows:

$$AUC_{mm,om} = \frac{1}{T}\sum_{t=1}^{T}(\widehat{AUC}_{t;mm,om}) \tag{7}$$

$$Bias_{mm,om} = \frac{1}{T}\sum_{t=1}^{T}(\widehat{AUC}^{n}_{t;mm,om} - \widehat{AUC}^{N}_{t;mm}) \tag{8}$$

$$MSE_{mm,om} = \frac{1}{T}\sum_{t=1}^{T}(\widehat{AUC}^{n}_{t;mm,om} - \widehat{AUC}^{N}_{t;mm})^2 \tag{9}$$

where $\widehat{AUC}_{t;mm,om}$ denotes the generic estimator $\widehat{AUC}$ based on the different combination of methods, when computed from the t-th Monte Carlo trial, and where the upper index identify the respective sample. Note that missing method, $mm \in \{CC, IPW, MI\}$ is always the first one and the optimism method is the second, $om \in \{SS, KF, LOO\}$. See details in appendix A. The $\widehat{AUC}^{N}_{t;mm}$ is the out-of-sample AUC obtained with particular missing method *mm*; this is the target, and it varies from trial to trial since the fitted model varies too. We decided to generate a large sample (*bigdata*), because we are interested in an estimate with good accuracy and precision. This idea, used in Iparragirre et al. (2019), was previously considered by other authors (Austin and Steyerberg, 2017; Hsu and Chen, 2016; Smith et al., 2014; Steyerberg et al., 2001; Yan, Tian and Liu, 2015). It is important to mention that for *bigdata* no missing scenario was created, the sample was always complete.

4.2. Results

The simulation results are reported in Tables 1-4, and graphically summarized in Figures 2-4. From these results it is seen that the apparent AUC (denoted by APP in Tables 1-4 and Figures 2-4) overestimates the target. This was expected, since the apparent AUC

measures the discriminatory capacity of the model on the very data that were used to fit the model. The overestimation issue is much more evident with a small sample size; with $n = 2000$ the data become almost fully representative of the target population, so the issue vanishes to some extent.

Table 1 shows the AUC optimism corrections for the full data, which are not available in practice but are interesting for comparison purposes. The LOO method always presents the smallest MSE, sometimes equalled by KF. Regarding the bias, the closest to zero is always presented by KF, followed by LOO and finally SS, with the latter methods tending to underestimate the target (negative bias).

Table 1. *Optimism corrections for the AUC and respective bias and MSE. No missing data scenario.*

		$n = 250$			$n = 500$			$n = 2000$		
prev	method	AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7645	0.0924	0.0109	0.7351	0.0465	0.0034	0.7180	0.0130	0.0005
	SS	0.6454	-0.0267	0.0090	0.6697	-0.0189	0.0039	0.6988	-0.0063	0.0008
	KF	0.6711	-0.0010	0.0063	0.6863	-0.0023	0.0023	0.7045	-0.0006	0.0004
	LOO	0.6528	-0.0193	0.0057	0.6756	-0.0130	0.0022	0.7025	-0.0026	0.0004
0.2	APP	0.7427	0.0572	0.0047	0.7284	0.0307	0.0016	0.7149	0.0071	0.0003
	SS	0.6642	-0.0214	0.0051	0.6873	-0.0104	0.0020	0.7035	-0.0044	0.0005
	KF	0.6850	-0.0005	0.0028	0.6986	0.0009	0.0010	0.7070	-0.0009	0.0003
	LOO	0.6740	-0.0115	0.0026	0.6931	-0.0045	0.0010	0.7058	-0.0020	0.0003
0.5	APP	0.7317	0.0380	0.0024	0.7223	0.0203	0.0009	0.7147	0.0056	0.0001
	SS	0.6794	-0.0144	0.0028	0.6960	-0.0059	0.0012	0.7075	-0.0016	0.0002
	KF	0.6920	-0.0018	0.0015	0.7021	0.0001	0.0006	0.7095	0.0004	0.0001
	LOO	0.6853	-0.0085	0.0015	0.6985	-0.0034	0.0006	0.7087	-0.0005	0.0001

Table 2 corresponds to MCAR scenario S1. The CC and IPW methodologies provide very similar estimates to each other with a worse performance than MI, which reports the smallest MSE in all cases. In addition, the APP value with CC and IPW is greater than APP with MI (larger positive bias) and the CC/IPW corrections (SS, KF, LOO) overcorrect optimism (lower values) compared to MI corrections, especially with small prevalences and sample sizes. These results are illustrated in Figure 1, for $n = 500$ and $prev = 0.2$. In Figure 1, the three blocks of boxplots indicate the results of CC (left), IPW(middle) and MI (rigth) estimators, respectively, and the first boxplot of each block corresponds to the out-of-sample AUC (defined in Section 4) which is the target. Labels such as CC-OOS, CC-APP, CC-SS, ... indicate the specific estimation method used within each group. Considering MI, LOO is the correction method that presents the smallest MSE and the bias closest to zero. See Figure 2 for relative results on the bias of the several optimism correction methods when applied to MI with an increasing sample size. Note that from Table 2 to Table 4, the lowest MSE is shown in bold and the lowest bias is underlined.

Table 2. *Optimism corrections for the AUC and respective bias and MSE. MCAR scenario S1.*

$n = 250$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.8058	0.1674	0.0330	0.8079	0.1708	0.0343	0.7689	0.1071	0.0142
	SS	0.6024	-0.0360	0.0209	0.6008	-0.0362	0.0215	0.6508	-0.0110	0.0083
	KF	0.6345	-0.0039	0.0155	0.6317	-0.0053	0.0163	0.6822	0.0204	0.0066
	LOO	0.5931	-0.0453	0.0171	0.5870	-0.0500	0.0188	0.6607	-0.0010	**0.0056**
0.2	APP	0.7688	0.1070	0.0140	0.7704	0.1100	0.0147	0.7469	0.0687	0.0063
	SS	0.6341	-0.0277	0.0090	0.6340	-0.0264	0.0088	0.6682	-0.0100	0.0048
	KF	0.6622	0.0004	0.0066	0.6609	0.0005	0.0069	0.6918	0.0136	0.0032
	LOO	0.6394	-0.0224	0.0063	0.6369	-0.0235	0.0066	0.6794	0.0012	**0.0029**
0.5	APP	0.7518	0.0749	0.0076	0.7528	0.0767	0.0080	0.7352	0.0474	0.0033
	SS	0.6533	-0.0235	0.0063	0.6518	-0.0243	0.0067	0.6831	-0.0047	0.0029
	KF	0.6747	-0.0022	0.0041	0.6724	-0.0037	0.0044	0.6960	0.0082	0.0016
	LOO	0.6622	-0.0147	0.0041	0.6605	-0.0156	0.0044	0.6893	0.0016	**0.0015**
$n = 500$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7604	0.0916	0.0106	0.7612	0.0929	0.0109	0.7417	0.0580	0.0048
	SS	0.6368	-0.0320	0.0101	0.6364	-0.0319	0.0101	0.6771	-0.0066	0.0037
	KF	0.6655	-0.0033	0.0059	0.6648	-0.0035	0.0060	0.6914	0.0078	0.0027
	LOO	0.6470	-0.0218	0.0054	0.6457	-0.0225	0.0055	0.6831	-0.0006	**0.0024**
0.2	APP	0.7402	0.0555	0.0044	0.7405	0.0562	0.0045	0.7294	0.0353	0.0021
	SS	0.6626	-0.0221	0.0048	0.6617	-0.0227	0.0050	0.6875	-0.0066	0.0019
	KF	0.6835	-0.0012	0.0025	0.6836	-0.0008	0.0026	0.6991	0.0050	0.0012
	LOO	0.6722	-0.0125	0.0024	0.6713	-0.0130	0.0025	0.6943	0.0002	**0.0012**
0.5	APP	0.7325	0.0395	0.0025	0.7327	0.0398	0.0025	0.7241	0.0245	0.0011
	SS	0.6803	-0.0127	0.0029	0.6792	-0.0136	0.0031	0.6942	-0.0053	0.0012
	KF	0.6929	-0.0002	0.0015	0.6924	-0.0005	0.0015	0.7043	0.0047	0.0007
	LOO	0.6862	-0.0069	0.0014	0.6856	-0.0073	0.0015	0.7005	0.0010	**0.0006**
$n = 2000$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7228	0.0236	0.0012	0.7227	0.0235	0.0012	0.7207	0.0168	0.0007
	SS	0.6863	-0.0129	0.0019	0.6860	-0.0132	0.0019	0.7016	-0.0023	0.0008
	KF	0.6962	-0.0031	0.0009	0.6960	-0.0032	0.0009	0.7073	0.0034	**0.0005**
	LOO	0.6921	-0.0071	0.0009	0.6918	-0.0074	0.0009	0.7055	0.0015	**0.0005**
0.2	APP	0.7185	0.0140	0.0006	0.7184	0.0139	0.0006	0.7154	0.0083	**0.0003**
	SS	0.6970	-0.0075	0.0010	0.6969	-0.0076	0.0010	0.7043	-0.0028	0.0005
	KF	0.7025	-0.0020	0.0005	0.7023	-0.0021	0.0005	0.7076	0.0006	**0.0003**
	LOO	0.7004	-0.0041	0.0005	0.7003	-0.0042	0.0005	0.7063	-0.0007	**0.0003**
0.5	APP	0.7161	0.0094	0.0004	0.7161	0.0095	0.00044	0.7143	0.0060	0.0002
	SS	0.7010	-0.0057	0.0006	0.7010	-0.0057	0.0006	0.7071	-0.0012	0.0003
	KF	0.7059	-0.0008	0.0003	0.7058	-0.0008	0.0003	0.7090	0.0007	**0.0002**
	LOO	0.7041	-0.0026	0.0003	0.7040	-0.0026	0.0003	0.7083	0.0000	**0.0002**

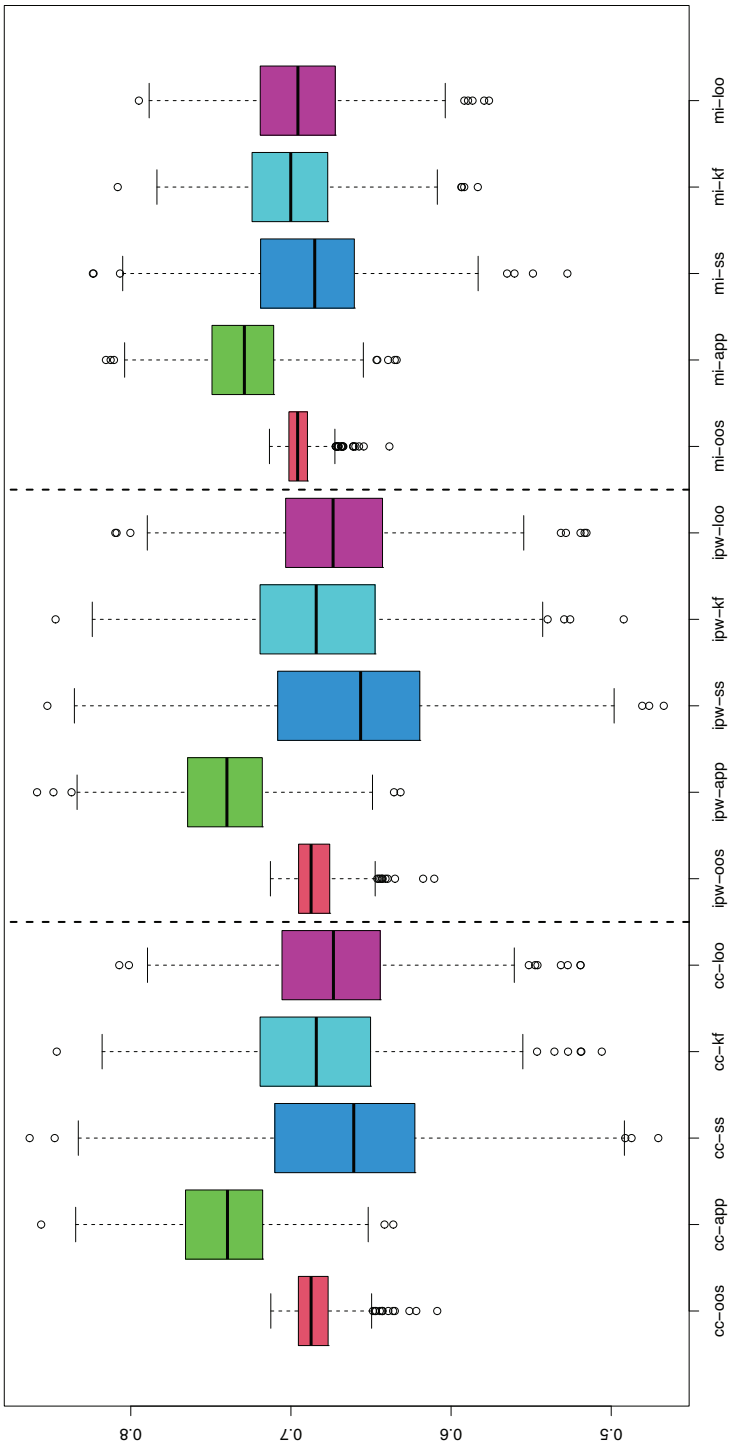

Figure 1. *Boxplot of AUC estimates with* $n = 500$ *and* $prev = 0.2$ *in the MCAR scenario S1 : CC left panel, IPW middle panel, and MI right panel.*

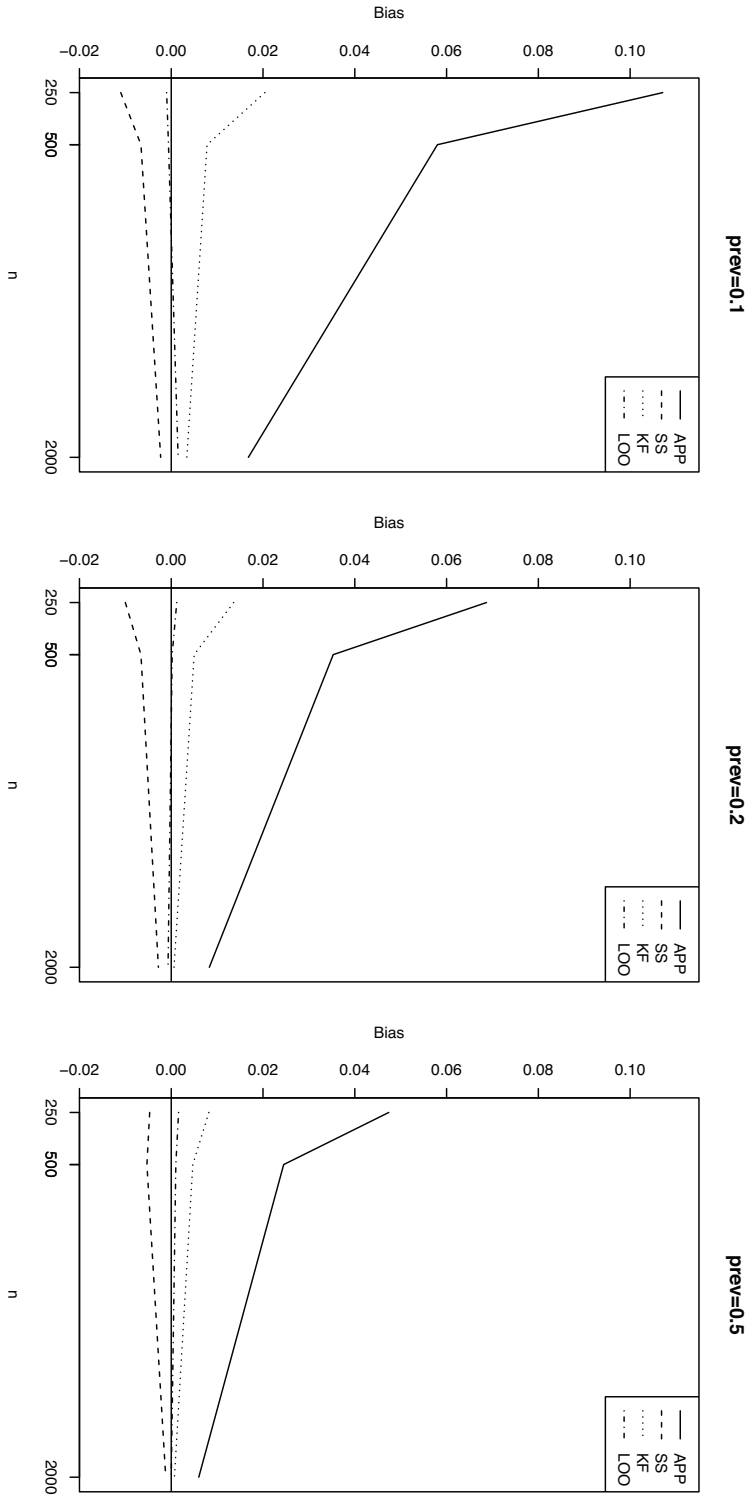

Figure 2. *Bias associated with each optimistic correction method with MI according to different sample sizes and prevalence in MCAR scenario S1.*

Table 3. *Optimism corrections for the AUC and respective bias and MSE. MAR scenario S2.*

$n = 250$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7972	0.1582	0.0306	0.8455	0.2123	0.0506	0.7679	0.1084	0.0142
	SS	0.5855	-0.0534	0.0225	0.6026	-0.0306	0.0264	0.6590	-0.0005	0.0077
	KF	0.6221	-0.0168	0.0161	0.6487	0.0154	0.0149	0.6786	0.0191	0.0061
	LOO	0.5807	-0.0582	0.0181	0.5912	-0.0420	0.0218	0.6582	-0.0013	**0.0053**
0.2	APP	0.7602	0.1006	0.0130	0.8108	0.1648	0.0309	0.7437	0.0663	0.0057
	SS	0.6189	-0.0408	0.0113	0.6114	-0.0345	0.0179	0.6691	-0.0083	0.0041
	KF	0.6434	-0.0162	0.0077	0.6525	0.0066	0.0085	0.6857	0.0082	0.0026
	LOO	0.6199	-0.0398	0.0083	0.6085	-0.0375	0.0150	0.6750	-0.0024	**0.0024**
0.5	APP	0.7478	0.0748	0.0077	0.7867	0.1313	0.0207	0.7367	0.0504	0.0036
	SS	0.6395	-0.0335	0.0078	0.6215	-0.0339	0.0154	0.6872	0.0009	0.0028
	KF	0.6640	-0.0090	0.0050	0.6442	-0.0113	0.0071	0.6976	0.0112	0.0017
	LOO	0.6492	-0.0238	0.0048	0.6221	-0.0334	0.0123	0.6911	0.0047	**0.0016**
$n = 500$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7554	0.0873	0.0101	0.7986	0.1441	0.0242	0.7415	0.0582	0.0046
	SS	0.6405	-0.0277	0.0090	0.6454	-0.0092	0.0129	0.6759	-0.0074	0.0034
	KF	0.6604	-0.0078	0.0054	0.6811	0.0265	0.0075	0.6926	0.0093	0.0024
	LOO	0.6400	-0.0282	0.0062	0.6364	-0.0182	0.0110	0.6826	-0.0006	**0.0022**
0.2	APP	0.7339	0.0525	0.0042	0.7745	0.1075	0.0142	0.7312	0.0367	0.0022
	SS	0.6511	-0.0303	0.0053	0.6542	-0.0128	0.0089	0.6907	-0.0039	0.0019
	KF	0.6743	-0.0071	0.0030	0.6804	0.0134	0.0050	0.7017	0.0072	**0.0012**
	LOO	0.6620	-0.0194	0.0030	0.6565	-0.0105	0.0075	0.6968	0.0023	**0.0012**
0.5	APP	0.7290	0.0377	0.0024	0.7575	0.0808	0.0083	0.7241	0.0255	0.0013
	SS	0.6698	-0.0215	0.0032	0.6619	-0.0147	0.0060	0.6945	-0.0041	0.0013
	KF	0.6867	-0.0046	0.0016	0.6755	-0.0011	0.0029	0.7038	0.0052	**0.0008**
	LOO	0.6778	-0.0135	0.0017	0.6639	-0.0128	0.0045	0.7005	0.0019	**0.0008**
$n = 2000$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7181	0.0191	0.0010	0.7443	0.0564	0.0045	0.7191	0.0155	0.0007
	SS	0.6808	-0.0182	0.0020	0.6780	-0.0099	0.0047	0.7001	-0.0034	0.0009
	KF	0.6903	-0.0087	0.0010	0.6985	0.0106	0.0021	0.7061	0.0025	**0.0005**
	LOO	0.6863	-0.0127	0.0011	0.6811	-0.0068	0.0030	0.7036	0.0001	**0.0005**
0.2	APP	0.7117	0.0082	0.0005	0.7313	0.0350	0.0021	0.7149	0.0084	0.0003
	SS	0.6883	-0.0152	0.0012	0.6833	-0.0129	0.0027	0.7035	-0.0030	0.0005
	KF	0.6958	-0.0077	0.0006	0.6999	0.0037	0.0011	0.7071	0.0006	**0.0003**
	LOO	0.6925	-0.0110	0.0006	0.6890	-0.0072	0.0015	0.7058	-0.0007	**0.0003**
0.5	APP	0.7111	0.0050	0.0003	0.7241	0.0246	0.0012	0.7139	0.0058	0.0002
	SS	0.6946	-0.0115	0.0008	0.6906	-0.0088	0.0016	0.7067	-0.0014	0.0003
	KF	0.6996	-0.0066	0.0004	0.6960	-0.0035	0.0007	0.7088	0.0007	**0.0002**
	LOO	0.6977	-0.0085	0.0004	0.6931	-0.0063	0.0010	0.7078	-0.0003	**0.0002**

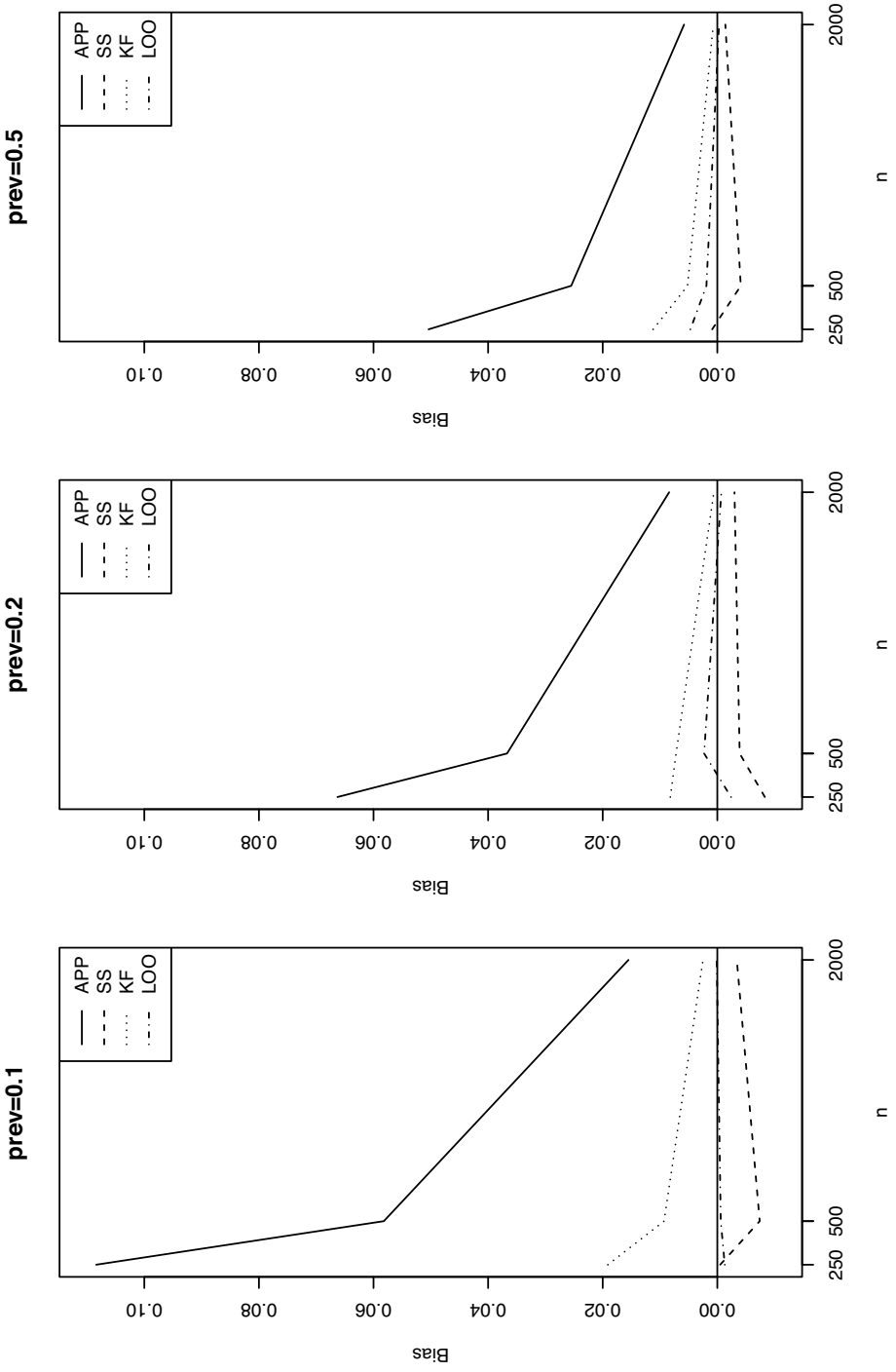

Figure 3. *Bias associated with each optimistic correction method with MI according to different sample sizes and prevalence in MAR scenario S2.*

Table 4. *Optimism corrections for the AUC and respective bias and MSE. MAR scenario S3.*

$n = 250$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7453	0.1232	0.0208	0.8009	0.1562	0.0294	0.7683	0.1037	0.0131
	SS	0.5708	-0.0513	0.0176	0.6039	-0.0408	0.0213	0.6528	-0.0117	0.0082
	KF	0.5910	-0.0311	0.0147	0.6317	-0.0130	0.0124	0.6763	0.0117	0.0058
	LOO	0.5612	-0.0608	0.0160	0.6023	-0.0424	0.0167	0.6590	-0.0056	**0.0051**
0.2	APP	0.7193	0.0806	0.0095	0.7757	0.1147	0.0164	0.7436	0.0623	0.0055
	SS	0.5934	-0.0453	0.0099	0.6207	-0.0403	0.0148	0.6691	-0.0122	0.0046
	KF	0.6162	-0.0225	0.0067	0.6412	-0.0198	0.0069	0.6866	0.0053	0.0031
	LOO	0.5970	-0.0416	0.0075	0.6276	-0.0334	0.0099	0.6752	-0.0062	**0.0029**
0.5	APP	0.7055	0.0600	0.0061	0.7612	0.1014	0.0135	0.7320	0.0433	0.0029
	SS	0.6034	-0.0421	0.0079	0.6191	-0.0407	0.0131	0.6800	-0.0087	0.0024
	KF	0.6209	-0.0247	0.0054	0.6220	-0.0378	0.0068	0.6924	0.0037	**0.0016**
	LOO	0.6055	-0.0401	0.0062	0.6221	-0.0377	0.0106	0.6855	-0.0032	**0.0016**
$n = 500$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7067	0.0578	0.0063	0.7646	0.0945	0.0116	0.7365	0.0515	0.0039
	SS	0.5997	-0.0491	0.0102	0.6431	-0.0270	0.0109	0.6702	-0.0148	0.0037
	KF	0.6190	-0.0298	0.0068	0.6647	-0.0054	0.0050	0.6861	0.0011	0.0027
	LOO	0.6022	-0.0466	0.0076	0.6472	-0.0229	0.0070	0.6763	-0.0087	**0.0022**
0.2	APP	0.6888	0.0285	0.0025	0.7484	0.0674	0.0064	0.7281	0.0332	0.0020
	SS	0.6137	-0.0466	0.0059	0.6534	-0.0275	0.0077	0.6889	-0.0061	0.0021
	KF	0.6287	-0.0316	0.0038	0.6661	-0.0149	0.0033	0.6980	0.0031	**0.0013**
	LOO	0.6196	-0.0407	0.0043	0.6628	-0.0181	0.0047	0.6928	-0.0021	**0.0013**
0.5	APP	0.6863	0.0200	0.0018	0.7380	0.0555	0.0048	0.7226	0.0230	0.0011
	SS	0.6254	-0.0409	0.0048	0.6518	-0.0307	0.0062	0.6948	-0.0049	0.0013
	KF	0.6407	-0.0256	0.0026	0.6549	-0.0276	0.0030	0.7024	0.0027	**0.0007**
	LOO	0.6329	-0.0334	0.0031	0.6611	-0.0215	0.0039	0.6989	-0.0008	**0.0007**
$n = 2000$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.6743	-0.0034	0.0008	0.7290	0.0320	0.0019	0.7186	0.0139	0.0006
	SS	0.6409	-0.0367	0.0030	0.6894	-0.0076	0.0026	0.6994	-0.0053	0.0008
	KF	0.6495	-0.0281	0.0017	0.6980	0.0010	0.0010	0.7051	0.0005	**0.0005**
	LOO	0.6450	-0.0326	0.0020	0.6925	-0.0044	0.0015	0.7032	-0.0014	**0.0004**
0.2	APP	0.6700	-0.0122	0.0006	0.7233	0.0213	0.0011	0.7163	0.0091	0.0003
	SS	0.6468	-0.0354	0.0022	0.6949	-0.0070	0.0018	0.7046	-0.0027	0.0004
	KF	0.6542	-0.0280	0.0013	0.6990	-0.0030	0.0007	0.7085	0.0013	**0.0002**
	LOO	0.6517	-0.0305	0.0015	0.6995	-0.0025	0.0009	0.7073	0.0001	**0.0002**
0.5	APP	0.6676	-0.0162	0.0006	0.7199	0.0179	0.0010	0.7143	0.0059	0.0002
	SS	0.6504	-0.0334	0.0018	0.6914	-0.0106	0.0018	0.7069	-0.0015	0.0003
	KF	0.6554	-0.0285	0.0012	0.6910	-0.0110	0.0008	0.7092	0.0008	**0.0002**
	LOO	0.6533	-0.0306	0.0013	0.6975	-0.0045	0.0010	0.7083	-0.0001	**0.0002**

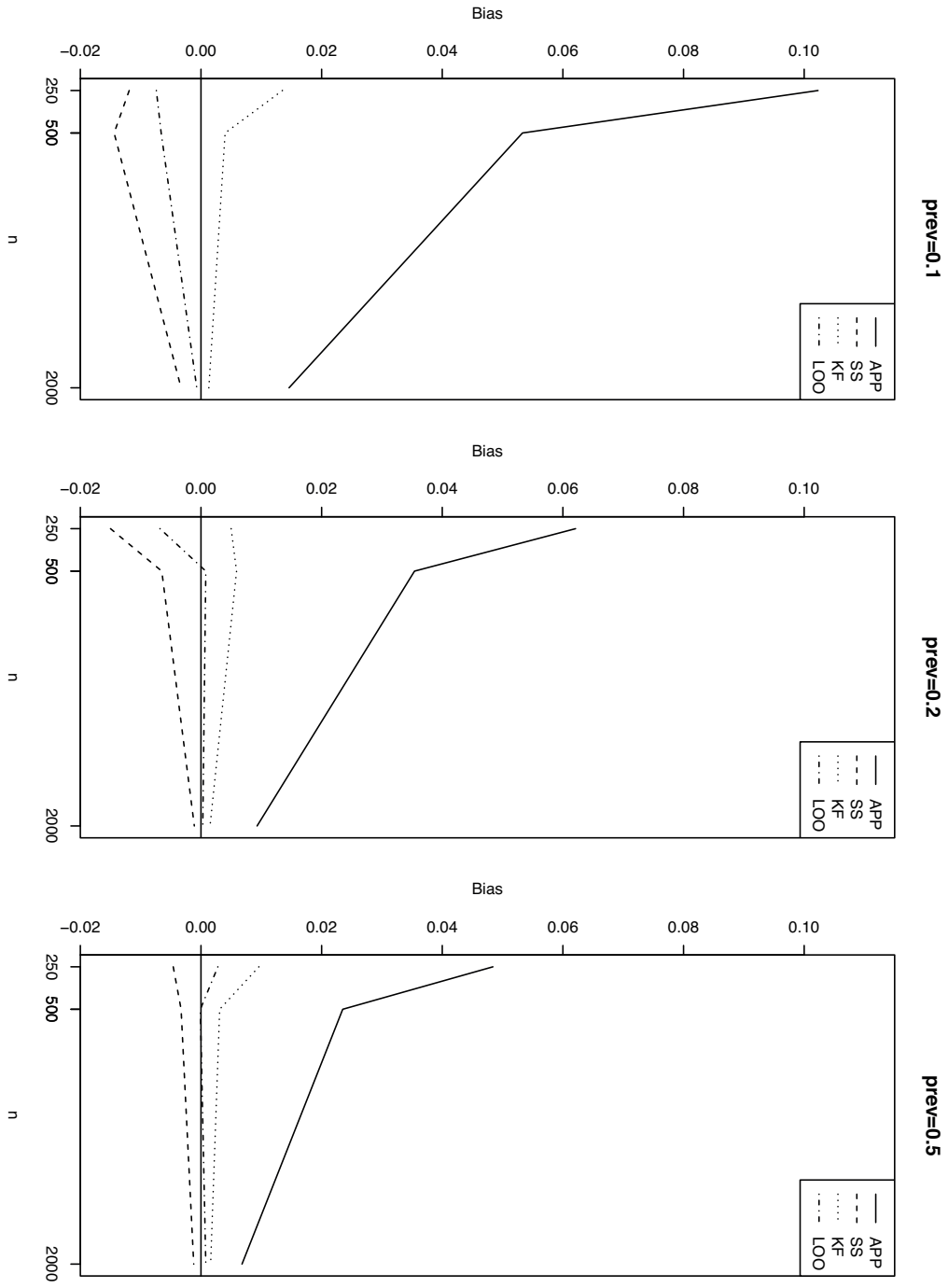

Figure 4. *Bias associated with each optimistic correction method with MI according to different sample sizes and prevalence in MAR scenario S3.*

In the MAR scenario where the loss of the covariate X_1 depends only on other covariates (Table 3, scenario S2) the MI estimators have the smallest MSE's again. In this scenario the CC and IPW estimators provide different values, which are larger than MI for the APP AUC and lower than MI for the SS, KF and LOO corrections. All methods seem to be consistent, but IPW estimators have larger MSE's due to the noise in the estimation of the weights Wi in equation (5). In fact, some simulations (not shown here) presented quite high values of the apparent AUC due to denominators very close to zero. Considering MI, LOO is the correction method that presents the smallest MSE, equalled by KF as n increases. Moreover, the bias of LOO is between that of KF and SS corrections, with the latter method tending to overcorrect for the optimism (see Figure 3).

Table 4 shows the results of the MAR scenario S3 where the missingness of the covariate X_1 also depends on the response. It is notable that the CC methodology is not consistent. Both the values of the APP method and those of the corrections are below the values obtained with MI, IPW or the corresponding values for the full data (Table 1), maintaining these differences for $n = 2000$. Again, IPW presents a MSE larger than MI, with estimated values of the AUC larger than MI for the APP estimator and lower than MI for the SS, KF and LOO corrections, which get closer as n increases. Considering MI, LOO is the correction method that presents the smallest MSE, similar to KF as n increases. As in previous scenarios, the bias of LOO is between that of KF and SS (see Figure 4). But in this case, when the prevalence is low, the bias closest to zero is generally obtained by KF.

Following the suggestion of a referee, we performed additional simulations for the bootstrap method. The bootstrap method aims to correct for optimism in model performance by comparing how well a model performs on resampled data versus the original data. Specifically, we applied the bootstrap method in combination with multiple imputation, since multiple imputation was the method for missing data that exhibiting the best performance. First, missing values were imputed, generating m completed datasets. Then, for each imputed dataset j, $1 \leq j \leq m$, the following steps were carried out:

1. Fit a logistic regression model to the imputed dataset j and estimate the apparent AUC, denoted $\widehat{\text{AUC}}_{\text{app}}^{(j)}$.

2. For $b = 1, \ldots, B$ (we used $B = 500$ bootstrap iterations):

   (a) Draw a bootstrap sample from the imputed dataset j (with replacement).

   (b) Fit a logistic regression model to the bootstrap sample and compute its apparent AUC, $\widehat{\text{AUC}}_{\text{boot}}^{(j,b)}$.

   (c) Apply the fitted model to the full imputed dataset j, and compute the AUC, $\widehat{\text{AUC}}_{\text{imp}}^{(j,b)}$.

3. Estimate the optimism for imputation j as the average difference:

$$\widehat{O}^{(j)} = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\mathrm{AUC}}_{\mathrm{boot}}^{(j,b)} - \widehat{\mathrm{AUC}}_{\mathrm{imp}}^{(j,b)}\right)$$

4. Compute the optimism-corrected AUC for imputation j:

$$\widehat{\mathrm{AUC}}_{\mathrm{corr}}^{(j)} = \widehat{\mathrm{AUC}}_{\mathrm{app}}^{(j)} - \widehat{O}^{(j)}$$

Finally, the overall corrected AUC estimate was obtained by averaging the corrected AUCs across the $m = 5$ imputations:

$$\widehat{\mathrm{AUC}}_{\text{mi-boot}} = \frac{1}{m}\sum_{j=1}^{m}\widehat{\mathrm{AUC}}_{\mathrm{corr}}^{(j)}$$

The results of the bootstrap method are reported in Table 11, Appendix B. From Table 11 one may see that MI-Bootstrap yields the lowest MSE in scenarios S1 and S3, especially for small samples, although it has a substantial bias. In scenario S2, the bootstrap method shows a MSE similar to that of the other methods, but the bias tends to be larger. It is also worth noting that the bootstrap method is computationally demanding, particularly when combined with multiple imputation to handle missing data.

In the context of complete data, Iparragirre et al. (2019) found that the bootstrap method (and K-fold cross-validation with replication) were the methods with longer waiting times, althought the computational effort was generally affordable. Our experience is in fully agreement with that. For instance, for the chronic lymphocytic leukemia dataset analyzed in Section 5.2, the computational times ranged from approximately 25 or 21 seconds for the most time-consuming method (MI+Boot and MI+LOO respectively) to 0.02 seconds for the fastest method (IPW+SS). In particular, corrections based on MI multiplied by a factor between five and six the computational times attached to IPW. This should be taken into account when planning intensive data analyses or simulation studies.

4.3. Discussion

In general, for all simulated scenarios and with all methodologies for missing data, the apparent AUC overestimates the true AUC, so this estimator is optimistic. This optimism tends to disappear when the sample size increases. Also, all correction methods reduce optimism, with both the MSE and the bias of the corrected estimators approaching zero as the sample size increases. These results are in line with the studies for complete data by Steyerberg et al. (2001); Airola et al. (2011) as well as with the results for missing data in Wahl et al. (2016).

In the MCAR scenario, CC and IPW methods perform similarly and are consistent, although MI is more efficient. These results are in agreement with Li et al. (2021) regarding APP AUC estimation and, to our knowledge, are shown here for the first time for

AUC correction methods with missing data. Considering MI, the best AUC correction method in terms of bias and MSE is LOO.

In the MAR scenarios, MI and IPW estimators showed consistency, with MI being more efficient. However, CC estimators were clearly inconsistent in estimating the AUC specially when the loss mechanism depends on the response variable. These results are in agreement with Li et al. (2021), who compared CC, IPW and MI in estimating the apparent AUC. In this study, the same behaviour was observed in the correction methods to estimate the AUC. Considering MI, the LOO correction method presented the lowest MSE, sometimes tied with KF. In terms of bias, the bias of LOO is usually between that of KF and SS corrections, with SS and sometimes LOO tending to overcorrect (negative bias), although to a lesser extent than in the complete data scenario. The pessimistic behaviour of LOO and SS with complete data is in agreement with Austin and Steyerberg (2017) or Iparragirre et al. (2019). To our knowledge, this is the first time that LOO correction of AUC has been studied in the context of missing data. Although LOO has been found pessimistic in complete data, our simulations indicate that such pessimism may be attenuated in missing data scenarios.

In all simulated MCAR and MAR scenarios, MI estimators of AUC presented the smallest MSEs relative to their corresponding CC and IPW estimators. Our results agree with, and extend to the context of the AUC corrections, the results by Li et al. (2021), who highlight the better performance of MI over CC and IPW estimators of the AUC, and the limitations of the IPW method when multiple covariates are missing. Recent studies by Wahl et al. (2016) and Mertens et al. (2020), in the context of correcting optimism with missing data, seem to assume this premise because they focus only on the use of the MI methodology.

The imputations have been performed taking the response variable into account. In general, the literature on MI recommends including the outcome variable in the imputation models (Von Hippel (2007); Little (1992)). However, the impact of the imputation model in both the estimation of the AUC and the correction of its optimism has been less investigated. In our study, the absolute differences between the AUC estimates with MI (Tables 2-4) and the respective estimates with complete data (Table 1) are not relevant, being below or around 0.01 in all cases. The order in which the imputation and correction methods are combined is also important. For this, two general strategies are possible. One possibility is to perform the optimism correction first and then to apply MI. An alternative approach is to impute first and the to correct for the optimism (MI-OM strategy). According to Wahl et al. (2016) the estimates obtained by MI-OM are optimistically biased . In our study, we performed simulations based on MI-OM idea, but we did not observe a significant increase in the estimations when the response variable was included in the imputation. We opted for initiating the simulation process with MI (or other missing data methodologies) in order to apply LOO correction, since prediction for an individual observation is not possible when some covariates are missing.

As a complement, we conducted additional simulations under the MCAR mechanism with missing probabilities of 0.2 and 0.8 (Appendix B). These confirmed that the

optimism of the apparent AUC increases with higher missingness, and that the MI+LOO combination remains the most accurate and robust across all tested scenarios.

As a summary, we can say that, according to our simulation results, the MI methodology with the LOO correction method is the best combination because it has the smallest MSE and an ignorable bias.

5. Application to real data

5.1. Schoolchildren of Viana do Castelo dataset

The methods for correcting AUC optimism in the presence of missing data were applied to a case study on obesity in municipal schools in Viana do Castelo, based on the dataset by Rodrigues, Bezerra and Saraiva (2008). This dataset includes 229 children from northern Portugal. A logistic regression model was used to predict the International Obesity Task Force ($IOTF_{10}$) indicator based on physical examinations (ABD), past obesity status ($IOTF_4$), and sex. The binary response variable ($IOTF_{10}$) equals 1 for the presence of overweight or obesity and 0 for its absence. The disease prevalence in the dataset is approximately 0.18. The variable 'sex' does not have any missingness. Physical examinations had a 5% missingness rate, and both past and current obesity status had a 6% missingness rate. According Little test (Little (1992)) (p-value = 0.454), missing data occur under MCAR mechanism. The coefficients of prediction models and and respective p-values for CC and IPW methods are reported in Table 5. In these models all variables are significant considering a significance level of 0.1. For MI it is not possible to define a single prediction model because MI combines results from multiple imputed datasets, yielding a model that is a summary of models.

Table 5. *Coefficients and p-values of CC and IPW prediction models.*

	CC		IPW	
variable	coeficient	p-value	coeficient	p-value
intercept	- 0.0825	0.9348	-0.0819	0.9333
sex	- 0.8759	0.0661	-0.8759	0.0579
$IOTF_4$	2.9606	5.91e-10	2.9605	1.63e-10
ABD	- 0.0861	0.0047	-0.0861	0.0035

To estimate the weights in the IPW methodology, we considered the variable 'sex', which is the only complete variable, and the following logistic model was obtained:

$$logit(P(R=1)) = 2.70805 + 0.05407 \times sex$$

and variable sex was not significant (p-value = 0.923), that is according to the results of Little test.

Table 6. *AUC values in obesity case study according different methods.*

method	CC	IPW	MI
APP	0.8977	0.8977	0.8998
SS	0.8819	0.8818	0.8920
KF	0.8701	0.8701	0.8846
LOO	0.8759	0.8759	0.8976

In the case of Viana do Castelo study, Table 6, CC and IPW present similar results for the apparent AUC and its several corrections, while MI yields slightly higher AUC values compared to the others. KF is the method that produces the lowest AUC values, particularly for CC and IPW. In this data set, the SS results do not correspond to what one could expect given the results of our simulation study, since this method presents a higher AUC than other correction methods. Actually, Viana do Castelo study differs from the simulation scenarios in Section 4 in that the optimism of the AUC is almost ignorable. Still, one can compare these results with the MCAR scenario with $prev = 0.2$ and $n = 250$, as it is the most comparable scenario. In this dataset, the missing data mechanism is MCAR, the data size is $n = 229$, and the prevalence is $prev = 0.18$.

5.2. Chronic lymphocytic leukemia dataset

In this section we consider the chronic lymphocytic leukemia dataset provided by the European Society for Blood and Marrow Transplantation, previously analyzed by Schetelig et al. (2017). This dataset includes 694 patients and the same variables used by Mertens et al. (2020) in their study on the Brier score. A logistic regression model was employed to predict each individual's disease status (Status) based on patient-related variables. The predictors include age, performance status at transplantation (perfstat), cytogenetic abnormalities (cyto), remission status (remstat), prior treatments (asct), donor characteristics (donor), sex match (sexm) (between donor and patient), and clinical conditions (cond). The variable perfstat has four levels: Karnofsky 100, Karnofsky 90, Karnofsky 80, and Karnofsky $<= 70$. The variable remstat has three levels: CR, PR, and SD/PD. The variable cyto has four levels: del17p, del11q, other, and no abnormality. The variable asct is dichotomous: no prior ASCT and prior ASCT. The variable donor has three levels: matched related, matched UD, and partially mismatched UD. The variable sexm has four levels: PATmaleDONmale, PATmaleDONfemale, PATfemaleDONmale, and PATfemaleDONfemale. Finally, the variable cond has three levels: NMA, RIC, and MAC.

The binary response variable, Status, takes the value 1 for diseased individuals and 0 for healthy individuals. The prevalence of disease is around 0.27. The response variable (disease status) and the covariates age, prior treatments and donor characteristics are complete (no missing data). On other hand, performance status has 9% of missingness, remission status has 6%, cytogenetic abnormalities has 25%, and there is a 1%

missingness each for sex match and clinical conditions. According to Little test (p-value = 5.59e-6), we reject the null hypothesis of missing completely at random. Violation of MCAR assumption brings concerns on the results provided on the application of the CC approach.

The coefficients of prediction models and respective p-values of CC and IPW are those in Table 7. For CC, only perfstat and PR are significant, considering a significance level of 0.1. For IPW, additionally to these variables, age, matched UD and PATmale-DONfemale are significant, considering a significance level of 0.1. The results provided by CC are reasonable, despite of its potential inconsistency (MCAR assuption was rejected). Since the AUC estimated from MI is averaged along five different models, these are not reported in Table 7.

Table 7. *Coefficients and p-values of CC and IPW prediction models.*

	CC		IPW	
variable	coeficient	p-value	coeficient	p-value
Intercept	-1.6133	0.0004	-1.6098	9.98e-06
age	0.2002	0.1864	0.2216	0.0634
Karnofsky 90	0.2221	0.4786	0.2148	0.3843
Karnofsky 80	1.1209	0.0017	1.1627	3.84e-05
Karnofsky $<= 70$	2.0004	0.0023	1.9881	0.0002
PR	-0.6992	0.0522	-0.6782	0.0178
SD/PD	0.0805	0.8301	0.1477	0.6170
del11q	-0.1201	0.7050	-0.0578	0.8198
other	-0.1860	0.5382	-0.1522	0.5287
no abnormality	0.2029	0.6013	0.1997	0.5223
prior ASCT	-0.0879	0.8461	-0.0086	0.9773
matched UD	0.3866	0.1461	0.3891	0.0625
partially mismatched UD	0.4276	0.2513	0.4399	0.1434
PATmaleDONfemale	0.4477	0.1297	0.4665	0.0466
PATfemaleDONmale	-0.2596	0.4538	-0.1730	0.5290
PATfemaleDONfemale	-0.2298	0.5589	-0.2613	0.4101
RIC	0.3063	0.2719	0.3050	0.1668
MAC	-0.0461	0.9065	0.0029	0.9924

To estimate the weights in the IPW methodology, we only use the complete variables. The final logistic models is the following:

$$logit(P(R=1)) = 0.6385 + 0.2535 \times age - 0.9365 \times asct + 0.3722 \times \text{donor\_matched UD} + 0.3118 \times \text{donor\_partially mismatched UD} - 0.3082 \times \text{Status}$$

In the logistic model of weighting, all variables are significant considering a significance level of 0.1, including the outcome (disease status).

Table 8. *AUC values in leukemia case study according to different methods.*

method	CC	IPW	MI
APP	0.7009	0.7080	0.6860
SS	0.5818	0.5772	0.6156
KF	0.6185	0.6350	0.6342
LOO	0.6201	0.6281	0.6318

The estimated AUCs are reported in Table 8. In this data set, the optimism of the apparent AUC is evident, as it consistently presents the highest values across all missing data methodologies. CC and IPW give higher values than MI in the APP and lower in the corrections. This is in agreement with the simulation results, see Table 4. Also, CC (at least KF and LOO) seems to move further away from MI and IPW. We compare these results with the MAR scenario S3 with $prev = 0.2$ and $n = 500$ because it is the most similar scenario, due to the similar sample size and prevalence, and the fact that the missing probability depends on the outcome. Among the correction methods, SS yields the lowest AUC values but tends to underestimate the true AUC, according to simulations. In this data set, for all missing data approaches, KF and LOO methods produce close results.

The proportion of missing data is a crucial factor influencing the performance of both missing data handling methods and optimism correction techniques. In our simulation study, we evaluated scenarios with varying levels of missingness and observed that higher missing data rates accentuate the differences between methods. Specifically, approaches such as CC and IPW tend to lose reliability as missingness increases, while MI combined with LOO consistently exhibits strong performance across all settings. The real data applications corroborate these findings. In the Viana do Castelo dataset, where the overall missingness was relatively low (under 10%), all methods produced similar AUC estimates, and the impact of optimism was negligible. In contrast, the leukemia dataset, which presented higher missingness levels (up to 25% in some variables), revealed more pronounced differences between methods. In particular, CC corrections often underestimated the AUC and showed less stable results compared to MI or IPW. These findings are consistent with our simulation results and emphasize the need to consider the proportion of missing data when choosing an appropriate estimation strategy.

6. Main conclusions

In this paper, corrections for the optimism of the AUC in the presence of missing data have been investigated. All methods successfully corrected the optimism in the AUC. An exception was CC which, as expected, failed to provide unbiased estimations when the missingness is not completely at random. Among the several methods being compared, LOO achieved the lowest MSE. This is an interesting finding, since LOO has been previously reported as too pessimistic with complete data. Correction methods performed

particularly well with MI. In practice, we recommend using the combination of MI with LOO because of its good relative performance. Importantly, LOO method for optimism correction of the AUC with missing data had not been considered in the related literature.

The two real data illustrations provided in this manuscript cover two different situations that may appear in practice. For the Viana do Castelo study, the optimism of the AUC is negligible, and all methods roughly report the same result. Missingness can be assumed to be completely at random in this case. However, in the leukemia study the optimism of the AUC is evident, and choosing one or another method matters. Specifically, it has been seen how CC introduces some bias in this case, probably related to the fact that the missing probability depends on the outcome. For the leukemia data, IPW or MI methods, with KF or LOO optimism corrections, provided close results.

In addition to the main results reported for a missing value probability of 0.5, further simulations were conducted with lower (0.2) and higher (0.8) missing proportions, as presented in Appendix B. These results confirmed that the benefit of applying optimism correction methods becomes more evident as missingness increases. The MI+LOO combination consistently provided the lowest bias and MSE, even under high missingness levels, reinforcing its robustness and practical relevance.

Key Conclusions and Recommendations

- The MI method proved to be the most effective approach for handling missing data, because it consistently yielded the lowest bias and MSE across nearly all simulated scenarios.
- To correct optimism in AUC estimates in the presence of missing data, we recommend using MI combined with LOO cross validation.
- The KF cross validation method also produced competitive results and may serve as a viable alternative, particularly when computational efficiency is a concern.
- In MCAR settings with large sample sizes and high prevalence, CC combined with LOO or KF can be considered a reasonable alternative to MI.
- However, CC is not recommended in MAR scenarios, as it tends to introduce substantial bias.

Acknowledgments

The authors thank one anonymous reviewer and an Executive Editor for comments and suggestions which have served to improve the paper.

A. Details of methods

In this section we present a structured details of implemented methods to correct the AUC optimism with missing data.

- **Complete case analyses (CC)**
  - Apparent in complete case analyses (CC-APP)
    * Only the complete observations are considered. There are the observations with no missing values, that is, the observations with $R_i = 1$.
    * Fit the logistic model to Y depending on covariables.
    * Predicted values of this model are computed.
    * The AUC_{cc-app} is calculated, according to equation (4).
  - Split-sample validation in complete case analyses (CC-SS)
    * Considered observations with no missing values.
    * The sample with complete cases is divided into two: *trainc* and *testc*.
    * Using *trainc* the logistic model is fit to Y depending on covariables.
    * The predicted values of this model are computed to *testc*.
    * The AUC_{cc-ss} is calculated, according to equation (4).
  - K-fold cross-validation in complete case analyses (CC-KF)
    * Only considering the complete observations, that is the observations with $R_i = 1$.
    * Divide the sample into K sets.
    * For $k = 1$ define the subset $test_k$ and the $train_k$ that is the sample without the $test_k$ set.
    * Fit a logistic model to Y depending by the $train_k$.
    * Computed the predicted values of this model to $testc_k$.
    * The AUC_k is calculated, according to equation (4).
    * Repeat this for each $k \in 2, 3, ..., K$.
    * The AUC_{cc-kf} is the mean of the AUC_k.
  - Leave-one-out cross-validation in complete case analyses (CC-LOO)
    * Considering the set with no missing values, that is, the observations with $R_i = 1$.
    * Define $trainc_j$ and $testc_j$.
    * Using $train_j$ fit a logistic model to Y.
    * The predicted values of this model are computed to $testc_j$.

  - The set of predictions, was construct with all of predictions of $test_j$
  - The AUC_{cc-loo} is calculated, according to equation (4).

- **Inverse probability weighting (IPW)**
  - Apparent in inverse probability weighting (IPW-APP)
    - Define the weighting of each no missing observation by $1/p_i$, where $p_i = E(R_i|Y,X) = E(R_i|Z)$.
    - The observations with no missing values, the observations with $R_i = 1$ are considered.
    - Fit a logistic model to Y depending on covariables considering all complete observations and the respective weight.
    - The predicted values of this data set are computed.
    - The $AUC_{ipw-app}$ is calculated, according to equation (5).
  - Split-sample validation in inverse probability weighting (IPW-SS)
    - Define the weighting of each no missing observation similar to IPW-APP.
    - The observations with no missing values, the observations with $R_i = 1$, are considered and divided into two sets: *trainc* and *testc*.
    - Fit a logistic model to Y depending on covariables considering the observations of *trainc* set and the respective weight.
    - The predicted values of *testc* with this model are computed.
    - The AUC_{ipw-ss} is calculated, according to equation (5).
  - K-fold cross-validation in inverse probability weighting (IPW-KF)
    - Define the weighting of each no missing observation.
    - Considering only the observations with no missing values and divide the sample into K sets.
    - For $k = 1$ define the subset $test_k$ and the $train_k$ that is the sample without the $test_k$ set.
    - Fit a logistic model to Y considering the observations of $trainc_k$ set and their weights.
    - The predicted values of $testc_k$ with this model and the respective AUC, AUC_k, are computed.
    - Repeat this for each $k \in 2,3,...,K$.
    - The AUC_{ipw-kf} is the mean of the AUC_k.
  - Leave-one-out cross-validation in inverse probability weighting (IPW-LOO)
    - Estimate the weighting model.

  - Define $train_j$ and $test_j$.
  - Fit a logistic model to Y considering the $train_j$ set and the respectively weights.
  - Computed the predict value of $test_j$.
  - Repeat this for all values of j.
  - The $AUC_{ipw-loo}$ is computed to the set of predictions.

- **Multiple imputation (MI)**
  - Apparent in multiple imputation (MI-APP)
    - The original data was imputed m times, getting m full sets $micdata_i$, $i \in 1,2,..,m$.
    - For each one, the logistic model to Y depending on covariables was considered.
    - The predicted values to each set are computed.
    - The AUC of each m sets is calculated.
    - The AUC_{mi-app} is calculated, averaging the previous AUC.
  - Split-sample validation in multiple imputation (MI-SS)
    - The original data was imputed m times, getting m full sets $micdata_i$, $i \in 1,2,..,m$.
    - Each of this sets was is divided into two: $mictrain_i$ and $mictest_i$.
    - For each one, the logistic model to Y depending on covariables of $mictrain_i$ set was considered.
    - The predicted values to each set $mictest_i$ are computed.
    - The AUC of each m *test* sets is calculated.
    - The AUC_{mi-ss} is calculated, averaging the previous AUC.
  - K-fold cross-validation in multiple imputation (MI-KF)
    - The data set was imputed m times, getting m full sets $cdat_j$, $j \in 1,2,..,m$.
    - Divide each full set $cdat_j$ into K sets.
    - For $k = 1$ define the subset $mictest_{j_k}$ and the $mictrain_{j_k}$ that is the j complete set without the $mictest_{j_k}$.
    - The logistic model to Y of $mictrain_{j_k}$ set was considered.
    - The predicted values to each set $mictest_{j_k}$ and respective AUC, AUC_{j_k}, are computed.
    - The AUC_k is calculated averaging the AUC_{j_k}.
    - Repeat this for all values $k \in 2,3,...,K$.
    - The AUC_{mi-kf} is the mean of the AUC_k.

- Leave-one-out cross-validation in multiple imputation (MI-LOO)
  - Imputed the data set by m multiple imputations, getting m full sets.
  - To each set, define $mictrain_j$ and $mictest_j$.
  - Using $mictrain_j$ fit a logistic model to Y.
  - The predicted values of this model are computed to $mictest_j$.
  - The set of predictions, was construct with all of predictions of $mictest_j$ and the respective AUC is computed.
  - The AUC_{mi-loo} is the mean of the last m AUC's.

B. Additional simulation results

Table 9. *Optimism corrections for the AUC and respective bias and MSE. MCAR scenario S1 with a probability of missing 0.2.*

$n = 250$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.7702	0.1099	0.0152	0.7708	0.1107	0.0154	0.7596	0.0922	0.0108
	SS	0.6283	-0.0320	0.0127	0.6277	-0.0325	0.0128	0.6407	-0.0266	0.0093
	KF	0.6563	-0.0040	0.0077	0.6546	-0.0055	0.0078	0.6671	-0.0003	0.0062
	LOO	0.6307	-0.0296	0.0082	0.6298	-0.0304	0.0083	0.6457	-0.0216	**0.0057**
0.2	APP	0.7469	0.0696	0.0068	0.7470	0.0698	0.0068	0.7406	0.0575	0.0048
	SS	0.6543	-0.0230	0.0068	0.6539	-0.0232	0.0069	0.6639	-0.0192	0.0047
	KF	0.6730	-0.0043	0.0042	0.6719	-0.0053	0.0043	0.6842	0.0011	0.0029
	LOO	0.6610	-0.0163	0.0039	0.6602	-0.0170	0.0039	0.6718	-0.0113	**0.0028**
0.5	APP	0.7382	0.0503	0.0037	0.7383	0.0505	0.0037	0.7324	0.0403	0.0026
	SS	0.6750	-0.0128	0.0032	0.6746	-0.0132	0.0032	0.6784	-0.0137	0.0027
	KF	0.6896	0.0017	0.0020	0.6891	0.0013	0.0020	0.6927	0.0006	0.0016
	LOO	0.6809	-0.0069	0.0019	0.6802	-0.0075	0.0019	0.6860	-0.0061	**0.0015**

Table 10. *Optimism corrections for the AUC and respective bias and MSE. MCAR scenario S1 with probability of missing 0.8.*

$n = 250$	method	CC			IPW			MI		
prev		AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
0.1	APP	0.9012	0.3212	0.1153	0.9082	0.3287	0.1197	0.7936	0.1594	0.0300
	SS	0.5472	-0.0328	0.0488	0.5524	-0.0271	0.0500	0.6866	0.0525	0.0125
	KF	0.5624	-0.0176	0.0364	0.5614	-0.0180	0.0344	0.7135	0.0794	0.0142
	LOO	0.4666	-0.1134	0.0625	0.4655	-0.1140	0.0640	0.6960	0.0619	**0.0114**
0.2	APP	0.8457	0.2261	0.0587	0.8547	0.2382	0.0644	0.7640	0.1049	0.0140
	SS	0.5673	-0.0523	0.0281	0.5642	-0.0523	0.0294	0.6940	0.0349	0.0074
	KF	0.6136	-0.0060	0.0210	0.6120	-0.0045	0.0226	0.7109	0.0518	0.0074
	LOO	0.5456	-0.0740	0.0248	0.5372	-0.0793	0.0284	0.7012	0.0421	**0.0063**
0.5	APP	0.8018	0.1626	0.0308	0.8130	0.1785	0.0363	0.7499	0.0772	0.0079
	SS	0.5900	-0.0492	0.0187	0.5865	-0.0480	0.0197	0.6992	0.0265	0.0044
	KF	0.6288	-0.0104	0.0138	0.6247	-0.0098	0.0155	0.7130	0.0403	0.0042
	LOO	0.5972	-0.0420	0.0137	0.5933	-0.0412	0.0145	0.7064	0.0337	**0.0038**

Table 11. *Optimism corrections for the AUC and corresponding bias and MSE using MI and bootstrap across three missingness scenarios with approximately 0.5 probability of missing.*

	n	250			500			2000		
S	*prev*	AUC	Bias	MSE	AUC	Bias	MSE	AUC	Bias	MSE
S1	0.1	0.6960	0.0365	0.0050	0.7009	0.0171	0.0021	0.7085	0.0044	0.0006
	0.2	0.6951	0.0174	0.0027	0.7049	0.0109	0.0011	0.7082	0.0014	0.0003
	0.5	0.7006	0.0137	0.0016	0.7042	0.0054	0.0007	0.7091	0.0011	0.0002
S2	0.1	0.6969	0.0404	0.0056	0.6990	0.0168	0.0021	0.7071	0.0035	0.0005
	0.2	0.6995	0.0229	0.0027	0.7014	0.0089	0.0013	0.7072	0.0002	0.0003
	0.5	0.7014	0.0149	0.0017	0.7055	0.0066	0.0008	0.7096	0.0016	0.0002
S3	0.1	0.6911	0.0263	0.0044	0.6952	0.0103	0.0018	0.7055	0.0012	0.0000
	0.2	0.6971	0.0164	0.0022	0.6997	0.0043	0.0009	0.70955	0.0026	0.0000
	0.5	0.6990	0.0109	0.0016	0.7055	0.0056	0.0008	0.7091	0.0009	0.0000

Funding

Work supported by the grants PID2020-118101GB-I00, Ministerio de Ciencia e Innovación (MCIN/ AEI /10.13039/501100011033) and PID2023-148811NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

Airola, A., T. Pahikkala, W. Waegeman, B. De Baets, and T. Salakoski (2011). An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis 55*(4), 1828–1844.

Austin, P. C. and E. W. Steyerberg (2017). Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical Methods in Medical Research 26*(2).

Carpenter, J. R. and M. Smuk (2021). Missing data: A statistical framework for practice. *Biometrical Journal 63*(5), 915–947.

Chen, X., A. T. K. Wan, and Y. Zhou (2015). Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association 110*(510), 723–741.

Cho, H., G. J. Matthews, and O. Harel (2019). Confidence intervals for the area under the receiver operating characteristic curve in the presence of ignorable missing data. *International Statistical Review 87*(1), 152-177.

Fan, J., S. Upadhye, and A. Worster (2006). Understanding receiver operating characteristic (ROC) curves. *CJEM 8*(1), 19-20.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters 27*(8), 861–874.

Garcia-Gutierrez, S., J. M. Quintana, A. Antón-Ladislao, M. S. Gallardo, E. Pulido, I. Rilo, E. Zubillaga, M. Morillas, J. J. Onaindia, N. Murga, R. Palenzuela, and J. G. Ruiz (2017). Creation and validation of the acute heart failure risk score: AHFRS. *Internal and Emergency Medicine 12*, 1197-1206.

Hsu, M.-J. and Y.-H. Chen (2016). Optimal linear combination of biomarkers for multicategory diagnosis. *Statistics in Medicine 35*(2), 202–213.

Iparragirre, A., I. Barrio, and M. X. Rodriguez-Alvarez (2019, 1). On the optimism correction of the area under the receiver operating characteristic curve in logistic prediction models. *SORT-Statistics and Operations Research Transactions 1*(1), 145–162.

Li, P., J. M. Taylor, D. E. Spratt, R. J. Karnes, and M. J. Schipper (2021). Evaluation of predictive model performance of an existing model in the presence of missing data. *Statistics in Medicine 40*(15), 3477–3498.

Little, R. J. A. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association 87*(420), 1227–1237.

Mertens, B. J. A., E. Banzato, and L. C. de Wreede (2020). Construction and assessment of prediction rules for binary outcome in the presence of missing predictor data using multiple imputation and cross-validation: Methodological approach and data-based evaluation. *Biometrical Journal 62*(3), 724–741.

Molenberghs, G. and M. G. Kenward (2007). *Missing Data on Clinical Studies*. Wiley.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Quintana, J. M., C. Esteban, A. Unzurrunzaga, S. Garcia-Gutierrez, N. Gonzalez, I. Lafuente, M. Bare, N. F. de Larrea, and S. Vidal (2014). Prognostic severity scores for patients with COPD exacerbations attending emergency departments. *The International Journal of Tuberculosis and Lung Disease 18*, 1415–1420.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Raghunathan, T. E., J. M. Lepkowski, J. V. Hoewyk, and P. W. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology 27*, 85–95.

Rodrigues, L., P. Bezerra, and L. Saraiva (2008). Morfologia e crescimento dos 6 aos 10 anos de idade em Viana do Castelo, Portugal. *Motricidade 4*.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. CRC press.

Schetelig, J., L. C. de Wreede, M. van Gelder, N. S. Andersen, C. Moreno, A. Vitek, M. Karas, M. Michallet, M. Machaczka, M. Gramatzki, D. Beelen, J. Finke, J. Delgado, L. Volin, J. Passweg, P. Dreger, A. Henseler, A. van Biezen, M. Bornäuser, S. O. Schönl, N. K. on behalf of the CLL subcommittee, and C. M. W. Party (2017). Risk factors for treatment failure after allogeneic transplantation of patients with CLL: a report from the European Society for Blood and Marrow Transplantation. *Bone Marrow Transplantation 52*, 552-560.

Smith, G. C. S., S. R. Seaman, A. M. Wood, P. Royston, and I. R. White (2014). Correcting for optimistic prediction in small data sets. *Statistical Methods in Medical Research 180*(3).

Steyerberg, E., S. Bleeker, H. Moll, D. Grobbee, and K. Moons (2003). Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology 56*, 441–7.

Steyerberg, E. W., F. E. Harrell, G. J. Borsboom, M. J. C. Eijkemans, Y. Vergouwe, and J. D. F. Habbema (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology 54 8*(8), 774–81.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research 16*(3), 219–242.

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software 45*(3), 1–67.

Von Hippel, P. T. (2007). Regression with missing ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology 37*(1), 83–117.

Wahl, S., A.-L. Boulesteix, A. Zierer, B. Thorand, and M. A. van de Wiel (2016). Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. *BMC Medical Research Methodology 16*.

Wishart, G., C. Bajdik, E. Dicks, E. Provenzano, M. K. Schmidt, M. Sherman, D. C. Greenberg, A. R. Green, K. A. Gelmon, V.-M. Kosma, J. E. Olson, M. W. Beckmann, R. Winqvist, S. S. Cross, G. Severi, D. Huntsman, K. Pylkäs, I. Ellis, T. O. Nielsen, G. Giles, C. Blomqvist, P. A. Fasching, F. J. Couch, E. Rakha, W. D. Foulkes, F. M. Blows, L. R. Bégin, L. J. van't Veer, M. Southey, H. Nevanlinna, A. Mannermaa, A. Cox, M. Cheang, L. Baglietto, C. Caldas, M. Garcia-Closas, and P. D. P. Pharoah (2012). PREDICT Plus: development and validation of a prognostic model for early breast cancer that includes HER2. *British Journal of Cancer 107*, 800–807.

Yan, L., L. Tian, and S. Liu (2015). Combining large number of weak biomarkers based on AUC. *Statistics in Medicine 34*, 3811–3830.

Zhu, J. and T. E. Raghunathan (2015). Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association 110*(511), 1112–1124.

SORT 49 (2) July-December 2025, 213-244 DOI: 10.57645/20.8080.02.28

On generalized Gower distance for mixed-type data: extensive simulation study and new software tools

Aurea Grané and Fabio Scielzo-Ortiz

Abstract

Data scientists address real-world problems using multivariate and heterogeneous datasets, characterized by multiple variables of different natures. Selecting a suitable distance function between units is crucial, as many statistical techniques and machine learning algorithms depend on this concept. Traditional distances, such as Euclidean or Manhattan, are unsuitable for mixed-type data, and although Gower distance was designed to handle this kind of data, it may lead to suboptimal results in the presence of outlying units or underlying correlation structure. In this work robust distances for mixed-type data are defined and explored, namely robust generalized Gower and robust related metric scaling. A new Python package is developed, which enables to compute these robust proposals as well as classical ones.

MSC: *62H30, 62-04.*

Keywords: *Distances, generalized Gower, multivariate heterogeneous data, outliers, robust Mahalanobis, related metric scaling.*

1. Introduction

Data scientists often face the challenge of clustering datasets of mixed-type, that is, datasets containing both numeric and categorical variables. A common approach is to start by computing classical Gower distance (Gower, 1971) between units, next to obtain a Euclidean configuration, for instance via metric multidimensional scaling (that is, Gower's 1966 principal coordinates, Borg and Groenen, 2005), and finally to apply partitioning algorithms like k-means or k-medians onto the principal coordinates of the

Authors' address: Statistics Deparment, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe, Spain. E-mails and ORCID code: A. Grané aurea.grane@uc3m.es (ORCID 0000-0003-0980-6409), F. Scielzo-Ortiz fscielzo@pa.uc3m.es.
Corresponding author: A. Grané.
Received: August 2024
Accepted: September 2025

units. Other possibilities skip the Euclidean configuration by directly applying clustering algorithms to classical Gower distance between units. This is the case for k-medoids (Kaufman and Rousseeuw, 1990) or hierarchical methods when the sample size allows it.

In such strategies, a key point is the selection of the metric, which should be able to incorporate the statistical characteristics of the data. For instance, the underlying correlation structure or outlying observations are issues that can distort the true proximity between units and that few metrics are able to consider. This may happen when using Gower distance, which is defined from Gower's similarity coefficient as the simple mean of three partial similarity indices computed from each variable type: a similarity associated with range-normalized Manhattan distance for numerical variables, Jaccard for binary variables and the simple matching coefficient for multiclass ones. Manhattan distance, like all Minkowski distances, implicitly assumes that variables are uncorrelated, and so does the Gower coefficient. Another problem is that the Manhattan distance is not robust to outlying units.

To overcome these drawbacks and inspired by Gower's work, the generalized Gower (G-Gower) distance was defined as the combination of three measures, conveniently standardized and fulfilling the Euclidean requirement (Gower and Legendre, 1986), for numerical, binary and multiclass variables (Grané, Salini and Verdolini, 2021). Indeed, G-Gower appears as a particular case when a more general technique, called related metric scaling (RelMS) (Cuadras and Fortiana, 1995; 1998), is used to tailor a metric. This technique allows combining several distance matrices computed on the same set of individuals into a single one. It has the additional property of discarding redundant information coming from different sources. When all distance matrices to be combined satisfy the Euclidean requirement, so does the final distance matrix (see Albarrán, Alonso and Grané, 2015; Grané and Romera, 2018 for the mathematical proofs).

In this paper, RelMS is used as a strategy to obtain flexible and robust distances for mixed-type data. Several proposals from the least to the greatest complexity are explored and evaluated in the context of clustering. They include three robust Mahalanobis proposals for numerical data, distances associated with Jaccard and Sokal-Michener similarity coefficients for binary data, and for multiclass variables Hamming distance is considered (which is the distance associated to the simple matching coefficient).

The performance of the new robust proposals is evaluated in six mixed-type datasets, four synthetic and two real, with underlying correlation structure and outlier contamination. In each case, k-medoids algorithm is applied to find the clusters, and comparisons with the performance of other classical metrics are provided in terms of classification rate and adjusted Rand index. A total of 34 distances are evaluated. Since some of the datasets are rather complex, metric multidimensional scaling is used to visualize and illustrate the difference between the true and assigned class of the units. A sensitivity analysis on the parameters involved in the robust estimation of the new proposals is provided for each dataset. A study of their computational cost for large and very large datasets can be found in Appendix B. Additionally, in Appendix A a Python package called robust_mixed_dist is presented.

The paper proceeds as follows. In Section 2 we revisit related metric scaling and present the generalized Gower distance as well as several robust proposals. Their performance in the context of clustering and the sensitivity analysis can be found in Section 3. Section 4 contains the main conclusions, and some guidelines on robust_mixed_dist and the study on computational cost can be found in the appendices.

2. Distance proposals for mixed-type data

In this section, a general procedure for combining distance matrices computed on the same set of units is revisited. It was used to obtain distance measures for mixed-type data in Albarrán et al. (2015), Grané and Romera (2018), Grané et al. (2021) and Boj and Grané (2024) in the context of metric multidimensional scaling and distance-based predictive models, where a robust Mahalanobis distance was used for numerical variables and for binary and multiclass data Jaccard and Hamming distances were considered, respectively. In this paper, we explore other robust proposals and provide an extensive simulation study of their performance in the context of clustering.

The strategy to construct a joint distance begins by splitting the dataset according to each variable type (numerical, binary and multiclass), next to compute different distance matrices for each variable type, and finally combine them via related metric scaling (Cuadras and Fortiana, 1995; 1998).

Let $\mathbf{X}$ be an $n \times p$ a data matrix corresponding to the measurements of p mixed-type variables $X_1, \ldots, X_p$ on a sample of n units, and consider sub-matrices $\mathbf{X}_k$ of size $n \times p_k$, $k = 1, 2, 3$, corresponding to each variable type, i.e., numeric, binary and multiclass, with $\sum_{k=1}^{3} p_k = p$. The distance measures considered are:

- Distances for numerical data: Euclidean (ℓ^2 distance), Manhattan (ℓ^1 distance), Canberra, Pearson (standardized ℓ^2 distance), Mahalanobis, robust Mahalanobis (with three variance estimators median absolute deviation, trimmed, winsorized),
- Distances for binary data: Associated with Jaccard coefficient (Jaccard, 1901) and with simple matching Sokal-Michener coefficient (Sokal and Michener, 1958).
- Distances for multiclass data: Hamming (associated to simple matching coefficient).

Most of the above distances are well-known to data scientists and their formulas are considered here. The previous list is not exhaustive, and other distances may be more appropriate depending on the context.

Regarding binary data, two similarity coefficients are considered, Jaccard and Sokal-Michener. It is worth noting that the Jaccard coefficient excludes double-zeros from the similarity assessment, whereas the Sokal-Michener coefficient takes them into account. Thus, the use of Jaccard coefficient is recommended when double-zeros are not informative. Otherwise, the Sokal-Michener coefficient is preferred (see Gower and Legendre,

1986 and Legendre and De Cáceres, 2013 for details and discussion). In any case, the general transformation given in Gower (1966) is considered to obtain a distance from a similarity coefficient. That is, consider the sub-matrix $\mathbf{X}_2$ corresponding to the measurements of p_2 binary variables on the sample of n units. The (squared) distance between units i,r is obtained as

$$\delta^2(\mathbf{x}_{2,i},\mathbf{x}_{2,r}) = s(\mathbf{x}_{2,i},\mathbf{x}_{2,i}) + s(\mathbf{x}_{2,r},\mathbf{x}_{2,r}) - 2\,s(\mathbf{x}_{2,i},\mathbf{x}_{2,r}), \tag{1}$$

where $\mathbf{x}_{2,i}$, $\mathbf{x}_{2,r}$ are $p_2 \times 1$ vectors containing the binary measurements for units i,r, respectively, and $s(\mathbf{x}_{2,i},\mathbf{x}_{2,r})$ is a given similarity coefficient between them.

Regarding numerical data, combinations including Euclidean, Manhattan, Pearson or Canberra distances are not recommended in the presence of an underlying correlation structure or outlying observations. In such cases, robust Mahalanobis proposals are preferred. A robust Mahalanobis distance is obtained by using a robust estimator for the covariance matrix in Mahalanobis distance formula.

In what follows, we focus on a procedure to obtain such a robust estimation, which consists of three steps: estimation of variances, estimation of Pearson's correlation coefficients, and estimation of covariances (see Gnanadesikan, 1997 for the details).

Consider the sub-matrix $\mathbf{X}_1$ corresponding to the measurements of p_1 numerical variables on the sample of n units. The (squared) robust Mahalanobis distance between units i,r is defined as:

$$\delta^2_{Maha}(\mathbf{x}_{1,i},\mathbf{x}_{1,r}) = (\mathbf{x}_{1,i}-\mathbf{x}_{1,r})'\,\mathbf{S}^{*-1}\,(\mathbf{x}_{1,i}-\mathbf{x}_{1,r}) \tag{2}$$

where $\mathbf{x}_{1,i}$ and $\mathbf{x}_{1,r}$ are $p_1 \times 1$ vectors containing the measurements for units i,r, respectively, and $\mathbf{S}^* = (s^*_{jk})_{1\le j,k\le p_1}$ is a robust estimation of the sample covariance matrix of $\mathbf{X}_1$.

In this paper we consider three methods for computing the s^*_{jk}'s, namely median absolute deviation (MAD), trimmed and winsorized. In any case, the first step of the procedure consists in selecting one of these three methods to estimate the variances of the numerical variables (that is, the diagonal elements of $\mathbf{S}^*$):

(1) MAD: $\hat{\sigma}^{*2}(X_j) = MAD(X_j)^2 = [Me\,(|x_{ij} - Me(X_j)| : i = 1,...,n)]^2$, where $Me(\cdot)$ stands for the median.

(2) Trimmed: $\hat{\sigma}^{*2}(X_j) = \hat{\sigma}^2(X_j^\alpha)$, where X_j^α is an α-trimmed version of X_j, that is,

$$X_j^\alpha = \left\{x_{ij} : i \in \{1,\ldots,n\}, x_{ij} \in [Q(\alpha/2,X_j), Q(1-\alpha/2,X_j)]\right. \quad ,$$

where $Q(z,X_j)$ is the $z \times 100$ quantile of X_j, $z \in [0,1]$.

(3) Winsorized: $\hat{\sigma}^{*2}(X_j) = \hat{\sigma}^2(X_j^\alpha)$, where X_j^α is an α-winsorized version of X_j, that is, $X_j^\alpha = \left\{h(x) : x \in X_j\right.$, where function h is defined as

$$h(x) = \begin{cases} a(\alpha), & \text{if } x \in A(\alpha), \\ b(\alpha), & \text{if } x \in B(\alpha), \\ x, & \text{if } x \in X_j \text{ and } x \notin A(\alpha), x \notin B(\alpha), \end{cases}$$

where $a(\alpha)$ is the value of X_j that is immediately greater than $Q(\alpha/2, X_j)$, $b(\alpha)$ is the value of X_j that is immediately lower than $Q(1-\alpha/2, X_j)$ and $A(\alpha) = \{x_{ij} : x_{ij} \leq Q(\alpha/2, X_j)\}$, $B(\alpha) = \{x_{ij} : x_{ij} \geq Q(1-\alpha/2, X_j)\}$.

In the second step of the procedure, a robust estimator of Pearson's correlation coefficient between two numerical variables is given. For each pair of variables X_j and X_k, a robust estimation of their Pearson's correlation coefficient is computed as follows:

$$r^*_{jk} = \frac{\hat{\sigma}^{*2}_+ - \hat{\sigma}^{*2}_-}{\hat{\sigma}^{*2}_+ + \hat{\sigma}^{*2}_-},$$

where $\hat{\sigma}^{*2}_+$ and $\hat{\sigma}^{*2}_-$ are robust estimators of the variances of $Z_j + Z_k$ and $Z_j - Z_k$, respectively, with $Z_j = X_j / \sqrt{\hat{\sigma}^{*2}(X_j)}$ and $Z_k = X_k / \sqrt{\hat{\sigma}^{*2}(X_k)}$. Note that the same method to estimate the variances selected in the first step must be used for $\hat{\sigma}^{*2}_+$ and $\hat{\sigma}^{*2}_-$.

In the final step of the procedure, the off-diagonal elements in $\mathbf{S}^*$ are obtained. For each pair of variables X_j and X_k, a robust estimation of their covariance is obtained as:

$$s^*_{jk} = r^*_{jk} \sqrt{\hat{\sigma}^{*2}(X_j)\, \hat{\sigma}^{*2}(X_k)}.$$

In the simulation study, parameter α in trimmed and winsorized methods was set equal to the true proportion of outlying units. Additionally, a sensitivity study on the effect of this parameter on the classification rate is given for each dataset.

Note that formula (2) relies on the fact that $\mathbf{S}^*$ is positive definite. In case $\mathbf{S}^* \geq 0$, then the inverse in formula (2) is substituted by the corresponding Moore-Penrose pseudo-inverse. In case $\mathbf{S}^*$ is not positive semi-definite, that is, in case negative eigenvalues exist, a shrinkage scheme can be applied to its elements to ensure positive definiteness. Some proposals can be found in Devlin et al. (1975). For instance, the Devlin algorithm to obtain positive definite $\mathbf{S}^*$ is based on the following transformation:

$$\mathcal{G}(\mathbf{S}^*) = (g(s^*_{jk}))_{\{j,k=1,\ldots,p\}},$$

where

$$g(s^*_{jk}) = \begin{cases} 0, & \text{if } |s^*_{jk}| \leq z(\varepsilon), \\ z^{-1}(z(s^*_{jk}) + \varepsilon), & \text{if } s^*_{jk} < -z(\varepsilon), \\ z^{-1}(z(s^*_{jk}) - \varepsilon), & \text{if } s^*_{jk} > z(\varepsilon), \end{cases}$$

where ε is a small positive number, for example, $\varepsilon = 0.05$, $z(x) = \arctan h(x) = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right)$, $z^{-1}(x) = \tan h(x)$ and $z(\varepsilon) = z(0.05) \approx 0.05$. The algorithm is applied recursively until $\mathcal{G}(\mathbf{S}^*)$ is positive definite.

In what follows we proceed to describe related metric scaling (RelMS), a multivariate technique introduced by (Cuadras and Fortiana, 1995, 1998), with the aim of combining several distance matrices computed on the same set of individuals in a single one. The method is based on the construction of a joint metric that satisfies several axioms related to the property of identifying and discarding redundant information (see

Albarrán et al., 2015; Grané and Romera, 2018). It is a very general method that allows to combine several sources of information, whenever a distance function can be measured between units. Although in Section 3 we explore it for the combination of numerical, binary and multiclass variables, it can also be applied to combine other kinds of data, such as functional data, time series, images, manifolds, compositional data, etc. Another possibility is to group variables according to different sources of information and combine the resulting distance matrices (Grané et al., 2022). Here we give the general description of the method.

Let $\mathbf{X}$ be an $n \times p$ data matrix corresponding to the measurements of p variables $X_1, \ldots, X_p$ on a sample of n units, and consider that the p variables can be grouped in m different types or sources of information.

1. Split matrix $\mathbf{X}$ into m sub-matrices $\mathbf{X}_k$ of size $n \times p_k$, $k = 1, \ldots, m$, regarding each variable type or source of information.

2. For each sub-matrix $\mathbf{X}_k$ consider a proper distance measure between units, according to the characteristics of the data, δ_k, and compute the corresponding matrix of squared pairwise distances conveniently standardized by its geometric variability (Cuadras and Fortiana, 1995), so that all matrices to be combined are commensurate, that is:
$$\mathbf{\Delta}_k = \frac{1}{V_{\mathbf{\Delta}_k}} \left(\delta_k^2(\mathbf{x}_{k,i}, \mathbf{x}_{k,r})\right)_{\{1 \leq i,r \leq n\}}, \tag{3}$$
where $\mathbf{x}_{k,i}, \mathbf{x}_{k,r}$ denote the i-th and r-th rows of matrix $\mathbf{X}_k$, respectively, and
$$V_{\mathbf{\Delta}_k} = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{r=1}^{n} \delta_k^2(\mathbf{x}_{k,i}, \mathbf{x}_{k,r}).$$

3. For each matrix $\mathbf{\Delta}_k$ compute the corresponding Gram matrix:
$$\mathbf{G}_k = -\frac{1}{2} \mathbf{H} \mathbf{\Delta}_k \mathbf{H},$$
where $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}'$ is the centering matrix, $\mathbf{I}$ is the identity matrix of size $n \times n$ and $\mathbf{1}$ is a $n \times 1$ vector of ones.

4. Check for *Euclideanarity*[1]: Each $\mathbf{G}_k$ must be positive definite. If this is not the case, several transformations can be applied to $\mathbf{\Delta}_k$ so that this requirement is fulfilled. In this paper the additive transformation is applied, but other possibilities may serve for this purpose (see Borg and Groenen, 1986; Gower and Legendre, 2005). For simplicity, we keep the same notation for $\mathbf{G}_k$'s, assuming that they satisfy the Euclidean requirement.

[1] The word "Euclideanarity" was coined by John Gower in Gower and Legendre (1986).

5. Combine all Gram matrices to get the Gram matrix of the joint metric as follows:

$$\mathbf{G} = \sum_{k=1}^{m} \mathbf{G}_k - \frac{1}{m} \sum_{k \neq l} \mathbf{G}_k^{1/2} \mathbf{G}_l^{1/2}, \tag{4}$$

where $\mathbf{G}_k^{1/2}$ is the square root of $\mathbf{G}_k$, which can be obtained through the singular value decomposition of $\mathbf{G}_k$.

6. The matrix of squared distances of the joint metric is obtained from $\mathbf{G}$ as follows:

$$\mathbf{\Delta} = \mathbf{g}\mathbf{1}' + \mathbf{1}\mathbf{g}' - 2\mathbf{G} = \left(\delta^2(\mathbf{x}_i, \mathbf{x}_r)\right)_{\{1 \leq i,r \leq n\}}, \tag{5}$$

where $\mathbf{g} = \text{diag}(\mathbf{G})$ is a $n \times 1$ vector containing the diagonal elements of $\mathbf{G}$, and $\mathbf{x}_i, \mathbf{x}_r$ denote the i-th and r-th rows of the data matrix $\mathbf{X}$, respectively.

7. Finally, the distance matrix of the joint metric is $\mathbf{D} = \left(\delta(\mathbf{x}_i, \mathbf{x}_r)\right)_{\{1 \leq i,r \leq n\}}$, that contains the square root of the elements of $\mathbf{\Delta}$.

The first addend of formula (4) mimics classical Gower distance by adding the three metrics, although here the addition is done through the matrices of square distances. The second addend is responsible of discarding redundant information coming from different sources. Note that RelMS can be computationally expensive for large sample sizes (see Appendix B). This is the reason why a simplified version of the above procedure was proposed, called generalized Gower distance (G-Gower).

Inspired by Gower's works, G-Gower (square) distance is defined as the linear combination of the matrices of squared pairwise distances, conveniently standardized by their corresponding geometric variability. That is,

$$\mathbf{\Delta}_{GG} = \sum_{k=1}^{m} \mathbf{\Delta}_k, \tag{6}$$

where each $\mathbf{\Delta}_k$ is defined as in (3) and fulfills the Euclidean requirement. Equivalently, (square) G-Gower can be obtained from the first addend of formula (4).

3. Empirical evaluation

In this section the performance of the distances presented in Section 2 is evaluated in the context of clustering and compared to those of classical Gower and Euclidean distance. The aim of the simulation study is to analyze their performance in the presence of underlying correlation structure and outlier contamination. The simulation involves four synthetic and two real datasets and k-medoids algorithm is used to obtain the partitioning.

In all cases, the true class is known for each unit. Thus, classification rate (proportion of units out of the total that are correctly classified) and adjusted Rand index

(ARI) (Hubert and Arabie, 1985; Rand, 1971) are used to evaluate the goodness of the clustering.

The Rand index was proposed by Rand (1971) as a clustering validation measure. However, as noted by Hubert and Arabie (1985) and Nguyen and Bailey (2009), in practice the Rand index frequently takes values in the $[0.5, 1]$ interval, its reference value (baseline value) can be high and not take a constant value. For these reasons the Rand index is most used in its adjusted version, known as the adjusted Rand index. Considering that there are n units and two partitions of them $\mathcal{C}_1 = \{C_{11}, \ldots, C_{1r}\}$, $\mathcal{C}_2 = \{C_{21}, \ldots, C_{2s}\}$ with r and s clusters, respectively, the adjusted Rand index is defined as

$$ARI = \frac{2\,(ab - cd)}{(a+c)(b+c) + (a+d)(b+d)},$$

where a is the number of pairs of units belonging to the same cluster in both partitions $\mathcal{C}_1$ and $\mathcal{C}_2$. That is, $i, j \in C_{1h}$ and $i, j \in C_{2u}$, for some $h = 1, \ldots, r$ and $u = 1, \ldots, s$; b is the number of pairs of units belonging to different clusters in partitions $\mathcal{C}_1$ and $\mathcal{C}_2$. That is, $i \in C_{1h_1}$ and $j \in C_{1h_2}$ for $h_1 \neq h_2$, $h_1, h_2 = 1, \ldots, r$, and also $i \in C_{2u_1}$ and $j \in C_{2u_2}$, for $u_1 \neq u_2$, $u_1, u_2 = 1, \ldots, s$; c is the number of pairs of units belonging to the same cluster in partition $\mathcal{C}_1$ but to different clusters in partition $\mathcal{C}_2$. That is, $i, j \in C_{1h}$ for $h = 1, \ldots, r$, but $i \in C_{2u_1}$ and $j \in C_{2u_2}$, for $u_1 \neq u_2$, $u_1, u_2 = 1, \ldots, s$; d is the number of pairs of units belonging to different clusters in partition $\mathcal{C}_1$ but to the same cluster in partition $\mathcal{C}_2$. That is, $i \in C_{1h_1}$ and $j \in C_{1h_2}$ for $h_1 \neq h_2$, $h_1, h_2 = 1, \ldots, r$, but $i, j \in C_{2u}$, for $u = 1, \ldots, s$.

The adjusted Rand index takes values in $[-0.5, 1]$ and the closer to one, the more similar the compared rankings are. In contrast, the closer to -0.5, the more different the compared rankings. In practice, one of the two cluster configurations (or two classifications) that the Rand index requires (adjusted or not), will be the one defined by a variable that is taken as a grouping response, and the other will be the one defined by a classification algorithm, such as k-medoids. Therefore, in practice it is necessary to have information about a categorical response variable to be able to implement the Rand index as a validation measure for clustering algorithms. In this scenario, the interpretation of the ARI is the following: the closer it is to 1, the more similar the classification made by the algorithm is to the real classification, and the closer it is to -0.5 , the less similar. In fact, in this context, an ARI close to zero indicates that the classification performed by the algorithm is similar to the one that would be obtained with a purely random classification procedure, and when it is negative it indicates that it is even worse.

A total of 34 distances are under study. In the case of G-Gower or RelMS, they are obtained as a combination of three distances, one for each type of data, following formulas (6) or (5), respectively. The distances considered are:

1. Generalized Gower composed by

    - Euclidean (ℓ^2 distance), Manhattan (ℓ^1 distance), Canberra, Pearson (standardized Euclidean), Mahalanobis, robust Mahalanobis (MAD, trimmed, winsozired) for numerical variables,

- Jaccard, Sokal-Michener similarity coefficients for binary variables, transformed to distances according to formula (1),
- Hamming for multiclass variables.

2. Related metric scaling composed by

   - Euclidean (ℓ^2 distance), Manhattan (ℓ^1 distance), Canberra, Pearson (standardized Euclidean), Mahalanobis, robust Mahalanobis (MAD, trimmed, winsozired) for numerical variables,
   - Jaccard, Sokal-Michener similarity coefficients for binary variables, transformed to distances according to formula (1),
   - Hamming for multiclass variables.

Additionally, Euclidean (ℓ^2 distance) and classical Gower distance are considered for comparison of results. Note that applying the Euclidean distance on raw mixed-type data is not recommended at all. The reason why we keep it is because it usually appears as the default distance in many software packages and we want to emphasize the consequences of using such a distance in a wrong context.

Tables 1–4 contain the classification rate and ARI mean values, computed on 100 runs for each scenario and distance considered. Additionally, Figures 1–4 contain metric MDS configurations corresponding to one of the 100 runs that help to illustrate the differences between the true and assigned class of the units. In the simulation study, results concerning the Euclidean distance are shown to illustrate the odd performance when using such a distance in mixed-type data.

3.1. Simulation study

Synthetic datasets of mixed-type data were generated with the `make_blobs` function from the `scikit-learn` Python library. Each dataset is composed by $p_1 = 4$ numerical, $p_2 = 2$ binary, $p_3 = 2$ multiclass variables measured on n units divided in k true classes. Different outlier patterns were added to each dataset.

1. Sample size: $n = 500$,
   Variables: Only 2 of them are informative and 6 with redundant information; Underlying correlation structure: Three pairs of numerical variables are highly correlated; Outliers: 10-12% contamination in three numerical variables.
   True classes: $k = 3$ (balanced).

2. Sample size: $n = 650$,
   Variables: 3 of them are informative and 5 with redundant information; Underlying correlation structure: Two pairs of numerical variables are highly correlated; Outliers: 10% contamination in two numerical variables,
   True classes: $k = 4$ (balanced).

Table 1. *Classification results for synthetic dataset 1.*

Distance	Classification rate	ARI
RelMS: Mahalanobis-Jaccard-Hamming	0.630	0.273666
G-Gower: Robust Mahalanobis trimmed-Jaccard-Hamming	0.628	0.278432
RelMS: Robust Mahalanobis winsorized-Jaccard-Hamming	0.628	0.277728
G-Gower: Mahalanobis-Jaccard-Hamming	0.628	0.268592
G-Gower: Robust Mahalanobis winsorized-Jaccard-Hamming	0.628	0.278432
RelMS: Canberra-Jaccard-Hamming	0.622	0.238612
G-Gower: Canberra-Jaccard-Hamming	0.620	0.230389
G-Gower: Canberra-Sokal-Hamming	0.620	0.218362
RelMS: Canberra-Sokal-Hamming	0.612	0.212984
G-Gower: Mahalanobis-Sokal-Hamming	0.608	0.222972
RelMS: Robust Mahalanobis trimmed-Jaccard-Hamming	0.598	0.255065
RelMS: Mahalanobis-Sokal-Hamming	0.590	0.195316
RelMS: Pearson-Jaccard-Hamming	0.588	0.190902
RelMS: Robust Mahalanobis winsorized-Sokal-Hamming	0.586	0.188601
RelMS: Robust Mahalanobis trimmed-Sokal-Hamming	0.584	0.185916
classical Gower	0.562	0.192146
G-Gower: Robust Mahalanobis MAD-Jaccard-Hamming	0.554	0.190833
G-Gower: Pearson-Jaccard-Hamming	0.552	0.190902
G-Gower: Pearson-Sokal-Hamming	0.540	0.220737
G-Gower: Robust Mahalanobis MAD-Sokal-Hamming	0.540	0.220737
RelMS: Pearson-Sokal-Hamming	0.538	0.202318
G-Gower: Euclidean-Sokal-Hamming	0.538	0.219598
RelMS: Robust Mahalanobis MAD-Sokal-Hamming	0.538	0.202318
G-Gower: Manhattan-Sokal-Hamming	0.534	0.216939
RelMS: Euclidean-Jaccard-Hamming	0.532	0.175767
G-Gower: Euclidean-Jaccard-Hamming	0.532	0.175767
RelMS: Robust Mahalanobis MAD-Jaccard-Hamming	0.530	0.170310
RelMS: Manhattan-Jaccard-Hamming	0.528	0.168527
G-Gower: Manhattan-Jaccard-Hamming	0.528	0.168527
RelMS: Manhattan-Sokal-Hamming	0.520	0.176607
RelMS: Euclidean-Sokal-Hamming	0.516	0.164035
G-Gower: Robust Mahalanobis trimmed-Sokal-Hamming	0.508	0.159580
G-Gower: Robust Mahalanobis winsorized-Sokal-Hamming	0.506	0.157313
Euclidean	0.350	0.000090

3. Sample size: $n = 600$,
   Variables: 6 of them are informative and 2 with redundant information; Underlying correlation structure: Three pairs of numerical variables are highly correlated;

Outliers: 7-8% contamination in three numerical variables.
True classes: $k = 4$ (unbalanced).

4. Sample size: $n = 600$,
   Variables: All informative; Uncorrelated and uncontaminated;
   True classes: $k = 4$ (balanced).

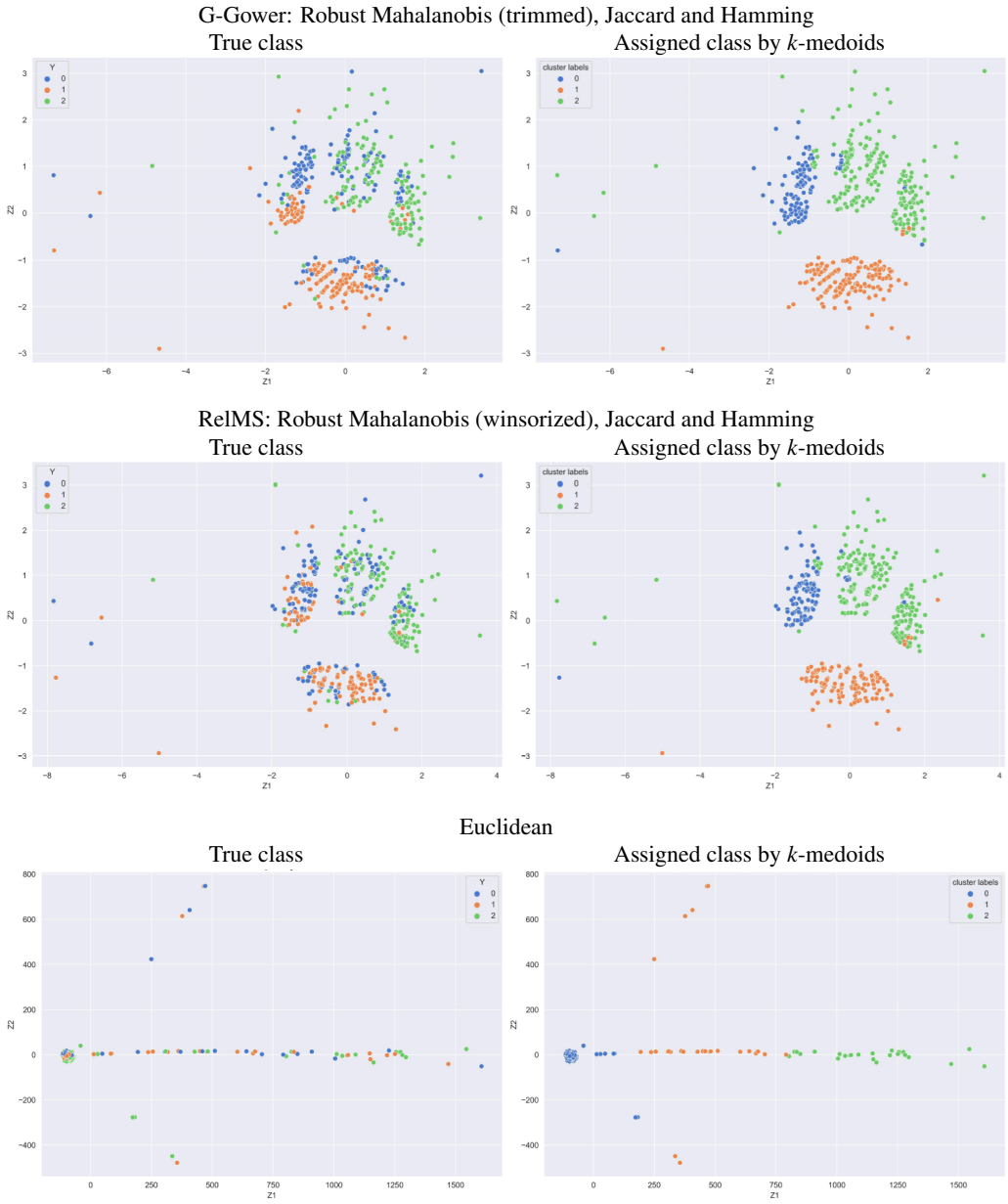

Figure 1. *Clustering visualization. Synthetic dataset 1.*

Table 2. *Classification results for synthetic dataset 2.*

Distance	Classification rate	ARI
G-Gower: Robust Mahalanobis MAD-Jaccard-Hamming	0.621538	0.329944
G-Gower: Mahalanobis-Sokal-Hamming	0.592308	0.324297
RelMS: Robust Mahalanobis MAD-Jaccard-Hamming	0.590769	0.260597
RelMS: Robust Mahalanobis MAD-Sokal-Hamming	0.567692	0.297498
RelMS: Mahalanobis-Sokal-Hamming	0.567692	0.228790
G-Gower: Canberra-Sokal-Hamming	0.561538	0.221440
RelMS: Robust Mahalanobis winsorized-Sokal-Hamming	0.558462	0.216499
RelMS: Robust Mahalanobis trimmed-Sokal-Hamming	0.558462	0.216499
RelMS: Pearson-Sokal-Hamming	0.552308	0.202442
G-Gower: Robust Mahalanobis trimmed-Sokal-Hamming	0.552308	0.214666
G-Gower: Robust Mahalanobis winsorized-Sokal-Hamming	0.552308	0.214415
G-Gower: Pearson-Sokal-Hamming	0.550769	0.215299
RelMS: Euclidean-Sokal-Hamming	0.547692	0.244515
G-Gower: Robust Mahalanobis MAD-Sokal-Hamming	0.546154	0.205740
G-Gower: Mahalanobis-Jaccard-Hamming	0.543077	0.226489
G-Gower: Robust Mahalanobis winsorized-Jaccard-Hamming	0.540000	0.222812
G-Gower: Manhattan-Sokal-Hamming	0.540000	0.232898
G-Gower: Euclidean-Sokal-Hamming	0.540000	0.232898
G-Gower: Robust Mahalanobis trimmed-Jaccard-Hamming	0.540000	0.222812
RelMS: Pearson-Jaccard-Hamming	0.533846	0.199402
RelMS: Mahalanobis-Jaccard-Hamming	0.526154	0.195419
G-Gower: Pearson-Jaccard-Hamming	0.524615	0.198994
G-Gower: Canberra-Jaccard-Hamming	0.524615	0.194698
RelMS: Euclidean-Jaccard-Hamming	0.503077	0.203928
G-Gower: Manhattan-Jaccard-Hamming	0.500000	0.193327
G-Gower: Euclidean-Jaccard-Hamming	0.500000	0.193327
RelMS: Canberra-Sokal-Hamming	0.496923	0.223695
RelMS: Manhattan-Jaccard-Hamming	0.495385	0.189415
RelMS: Manhattan-Sokal-Hamming	0.492308	0.191534
RelMS: Robust Mahalanobis trimmed-Jaccard-Hamming	0.481538	0.155862
RelMS: Robust Mahalanobis winsorized-Jaccard-Hamming	0.478462	0.153906
RelMS: Canberra-Jaccard-Hamming	0.460000	0.159691
classical Gower	0.423077	0.125406
Euclidean	0.263077	0.000138

Table 1 contains the classification rate and ARI values for k-medoids algorithm with $k = 3$, concerning synthetic dataset 1. We observe that G-Gower with robust Mahalanobis (trimmed), Jaccard and Hamming reaches a classification rate of 62.3% and an

ARI value of 0.27. Similar values are reached by RelMS with robust Mahalanobis (winsorized), Jaccard and Hamming. These classification rates (and ARI values) are higher than those obtained by classical Gower or Euclidean distance, for which values of 56.2% (ARI 0.19) and 35.0% (ARI $< 10^{-4}$) are attained, respectively.

In Figure 1 metric MDS maps are used to illustrate the k-medoids classification ($k = 3$), for synthetic dataset 1, using G-Gower with robust Mahalanobis (trimmed), Jaccard and Hamming, RelMS with robust Mahalanobis (winsorized), Jaccard and Hamming and Euclidean distance. Units in left panels are colored according to their true class and in right panels, according to their assigned class. The clustering in G-Gower and RelMS panels can be considered rather acceptable, since around half of the units in class 0 are not well identified. On the other hand, the clustering with Euclidean distance is disappointing, where the configuration appears completely distorted due to outlying observations and the underlying correlation structure that this distance is not able to incorporate.

Table 2 contains the classification rate and ARI values for k-medoids algorithm with $k = 4$, regarding synthetic dataset 2. In this case, G-Gower with robust Mahalanobis (MAD), Jaccard and Hamming reaches a classification rate of 62.2% and an ARI value of 0.33, and a 59.1% rate and 0.26 ARI are attained by RelMS with robust Mahalanobis (MAD), Jaccard and Hamming. These classification rates (and ARI values) are higher than those obtained by classical Gower or Euclidean distance, whose values are located at the end of the ranking.

In Figure 2 metric MDS maps are used to illustrate the k-medoids classification ($k = 4$), for synthetic dataset 2, using G-Gower with robust Mahalanobis (MAD), Jaccard and Hamming, RelMS with robust Mahalanobis (MAD), Jaccard and Hamming and classical Gower. Units in left panels are colored according to their true class and in right panels, according to their assigned class. The clustering in G-Gower panels can be considered rather acceptable, since most of the units in classes 0 and 1 are well identified. On the other hand, the clustering with classical Gower is not good since this distance is not able to incorporate the underlying correlation structure as well as the presence of outlying units. Once more, the clustering with Euclidean distance is disappointing.

Table 3 contains the classification rate and ARI values for k-medoids algorithm with $k = 4$, concerning synthetic dataset 3. We observe that G-Gower with robust Mahalanobis (MAD), Sokal and Hamming reaches a classification rate of 88.3% and an ARI value of 0.73. On the other hand, classification rates for Euclidean and classical Gower are of 44.17% (ARI 0.29) and 73.50% (ARI 0.26), respectively.

In Figure 3 metric MDS maps are used to illustrate the k-medoids classification ($k = 4$), for synthetic dataset 3, using G-Gower with robust Mahalanobis MAD, Sokal and Hamming and classical Gower. Units in left panels are colored according to their true class and in right panels, according to their assigned class. In G-Gower panels we can observe that the four clusters are very similar to the true ones. However, this is not the case for classical Gower, where most of the units in class 0 and half of the units in class 3 are not well identified.

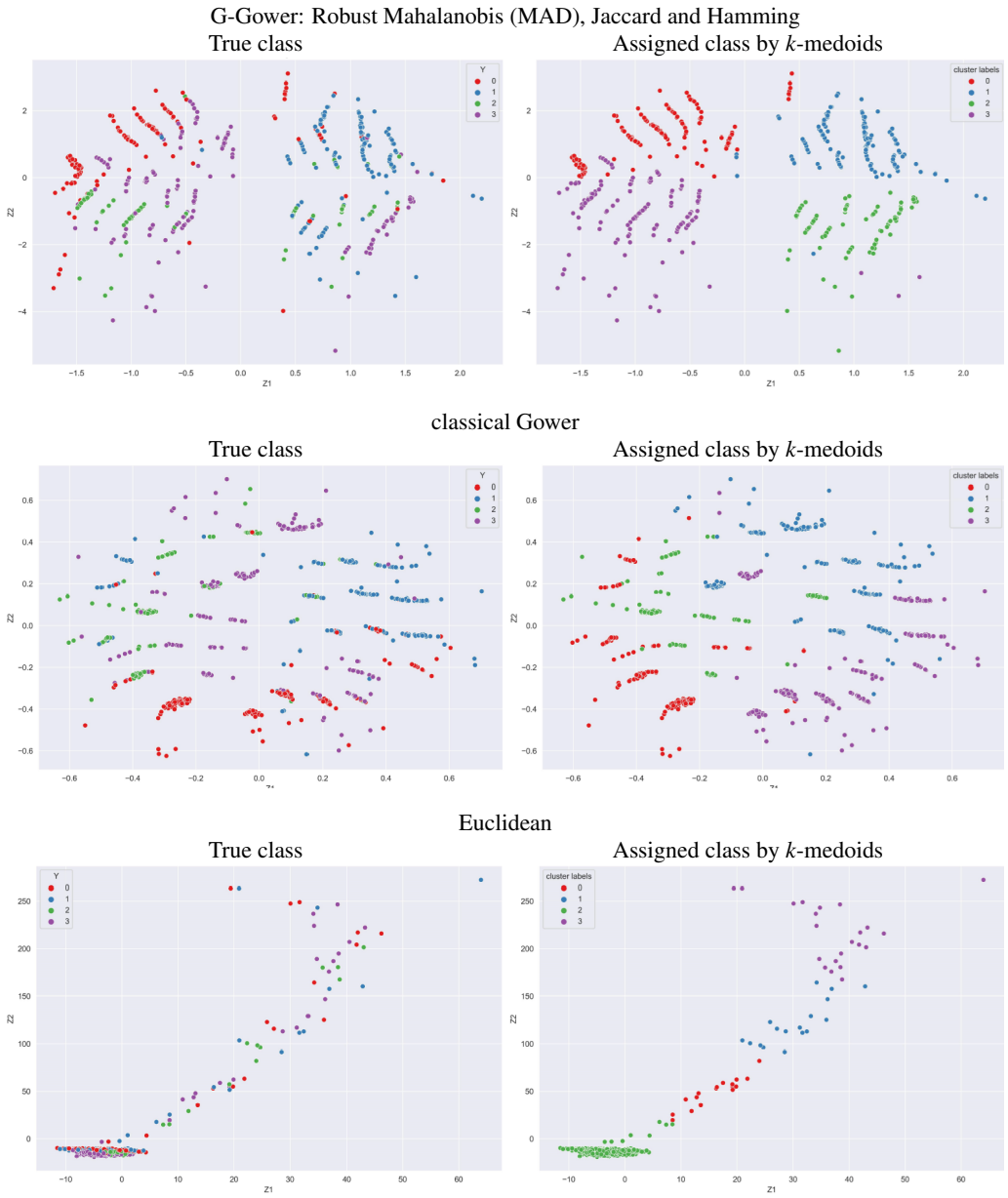

Figure 2. *Clustering visualization. Synthetic dataset 2.*

Table 3. *Classification results for synthetic dataset 3.*

Distance	Classification rate	ARI
G-Gower: Robust Mahalanobis MAD-Sokal-Hamming	0.883333	0.726101
G-Gower: Pearson-Sokal-Hamming	0.881667	0.723303
G-Gower: Canberra-Jaccard-Hamming	0.880000	0.704662
G-Gower: Canberra-Sokal-Hamming	0.880000	0.703744
RelMS: Canberra-Sokal-Hamming	0.866667	0.683853
RelMS: Canberra-Jaccard-Hamming	0.850000	0.657553
RelMS: Robust Mahalanobis MAD-Sokal-Hamming	0.731667	0.613787
RelMS: Pearson-Sokal-Hamming	0.730000	0.611196
RelMS: Euclidean-Sokal-Hamming	0.728333	0.615708
RelMS: Manhattan-Sokal-Hamming	0.728333	0.615395
G-Gower: Manhattan-Sokal-Hamming	0.723333	0.597885
G-Gower: Euclidean-Sokal-Hamming	0.710000	0.564313
G-Gower: Manhattan-Jaccard-Hamming	0.708333	0.559781
G-Gower: Euclidean-Jaccard-Hamming	0.708333	0.558163
RelMS: Manhattan-Jaccard-Hamming	0.705000	0.556431
RelMS: Euclidean-Jaccard-Hamming	0.700000	0.541697
G-Gower: Robust Mahalanobis winsorized-Jaccard-Hamming	0.690000	0.587217
G-Gower: Robust Mahalanobis MAD-Jaccard-Hamming	0.686667	0.589241
RelMS: Robust Mahalanobis MAD-Jaccard-Hamming	0.686667	0.589241
RelMS: Robust Mahalanobis winsorized-Jaccard-Hamming	0.686667	0.587224
G-Gower: Robust Mahalanobis trimmed-Jaccard-Hamming	0.686667	0.587224
RelMS: Mahalanobis-Jaccard-Hamming	0.686667	0.587224
G-Gower: Mahalanobis-Jaccard-Hamming	0.685000	0.576739
G-Gower: Pearson-Jaccard-Hamming	0.685000	0.582427
RelMS: Robust Mahalanobis trimmed-Jaccard-Hamming	0.683333	0.578341
RelMS: Pearson-Jaccard-Hamming	0.671667	0.547123
classical Gower	0.651667	0.292456
G-Gower: Robust Mahalanobis trimmed-Sokal-Hamming	0.605000	0.490818
RelMS: Robust Mahalanobis trimmed-Sokal-Hamming	0.605000	0.491105
G-Gower: Robust Mahalanobis winsorized-Sokal-Hamming	0.591667	0.486762
RelMS: Robust Mahalanobis winsorized-Sokal-Hamming	0.586667	0.484619
RelMS: Mahalanobis-Sokal-Hamming	0.583333	0.484672
G-Gower: Mahalanobis-Sokal-Hamming	0.581667	0.486274
Euclidean	0.533333	0.255522

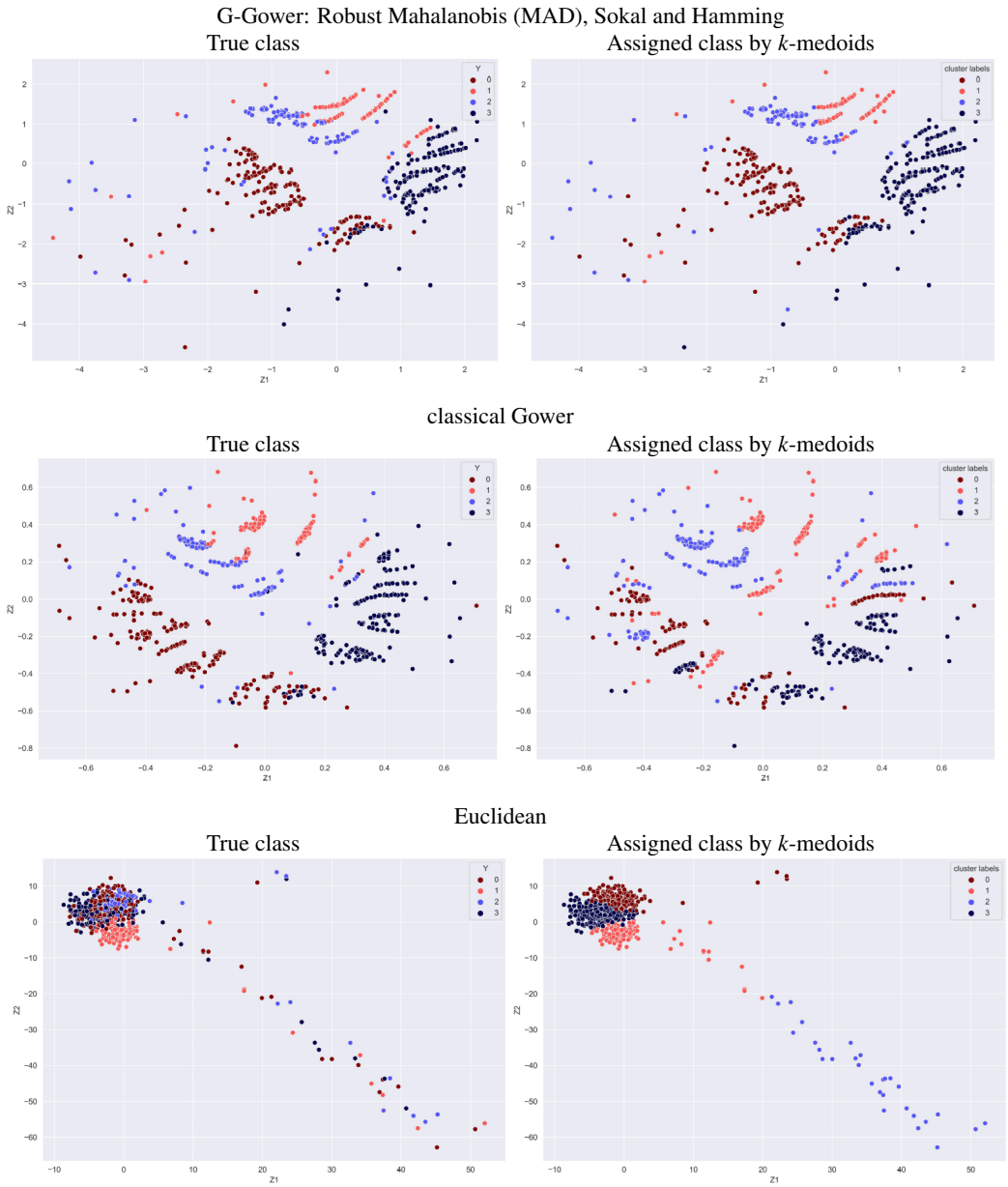

Figure 3. *Clustering visualization. Synthetic dataset 3.*

Table 4 contains the classification rate and ARI values for k-medoids algorithm with $k = 4$, concerning synthetic dataset 4, where variables are uncorrelated and uncontaminated. As expected, classification rates are rather low and most ARI values are close to zero, since data in the four groups lack underlying correlation structure (all variables are uncorrelated and either normally or uniformly distributed). In this case, the best performance is achieved by Euclidean distance, with a classification rate of 55.33% and

an ARI value of 0.1852, followed by G-Gower with Canberra distance for numerical variables. The performance of these methods is illustrated in Figure 4 with metric MDS maps.

Table 4. *Classification results for synthetic dataset 4.*

Distance	Classification rate	ARI
Euclidean	0.5533	0.1852
G-Gower: Canberra-Sokal-Hamming	0.4900	0.1652
RelMS: Canberra-Jaccard-Hamming	0.4750	0.1278
classical Gower	0.4683	0.1415
RelMS: Pearson-Jaccard-Hamming	0.4300	0.0905
RelMS: Euclidean-Sokal-Hamming	0.4300	0.1140
RelMS: Euclidean-Jaccard-Hamming	0.4283	0.0896
G-Gower: Manhattan-Jaccard-Hamming	0.4283	0.0918
G-Gower: Pearson-Jaccard-Hamming	0.4283	0.0898
G-Gower: Euclidean-Jaccard-Hamming	0.4267	0.0901
RelMS: Canberra-Sokal-Hamming	0.4267	0.1213
RelMS: Pearson-Sokal-Hamming	0.4217	0.1077
RelMS: Robust Mahalanobis mad-Sokal-Hamming	0.4200	0.1039
RelMS: Robust Mahalanobis winsorized-Jaccard-Hamming	0.4200	0.0841
RelMS: Manhattan-Jaccard-Hamming	0.4200	0.0826
RelMS: Robust Mahalanobis trimmed-Jaccard-Hamming	0.4200	0.0841
RelMS: Mahalanobis-Jaccard-Hamming	0.4200	0.0837
G-Gower: Mahalanobis-Jaccard-Hamming	0.4200	0.0845
G-Gower: Robust Mahalanobis winsorized-Jaccard-Hamming	0.4183	0.0844
G-Gower: Robust Mahalanobis trimmed-Jaccard-Hamming	0.4183	0.0845
RelMS: Mahalanobis-Sokal-Hamming	0.4150	0.1022
RelMS: Robust Mahalanobis mad-Jaccard-Hamming	0.4150	0.0808
RelMS: Robust Mahalanobis trimmed-Sokal-Hamming	0.4150	0.1022
RelMS: Robust Mahalanobis winsorized-Sokal-Hamming	0.4150	0.1022
G-Gower: Robust Mahalanobis mad-Jaccard-Hamming	0.4133	0.0816
G-Gower: Euclidean-Sokal-Hamming	0.3967	0.0841
G-Gower: Pearson-Sokal-Hamming	0.3917	0.0807
RelMS: Manhattan-Sokal-Hamming	0.3883	0.0928
G-Gower: Canberra-Jaccard-Hamming	0.3867	0.0528
G-Gower: Manhattan-Sokal-Hamming	0.3817	0.0874
G-Gower: Mahalanobis-Sokal-Hamming	0.3800	0.0882
G-Gower: Robust Mahalanobis trimmed-Sokal-Hamming	0.3800	0.0906
G-Gower: Robust Mahalanobis winsorized-Sokal-Hamming	0.3783	0.0887
G-Gower: Robust Mahalanobis mad-Sokal-Hamming	0.3783	0.0921

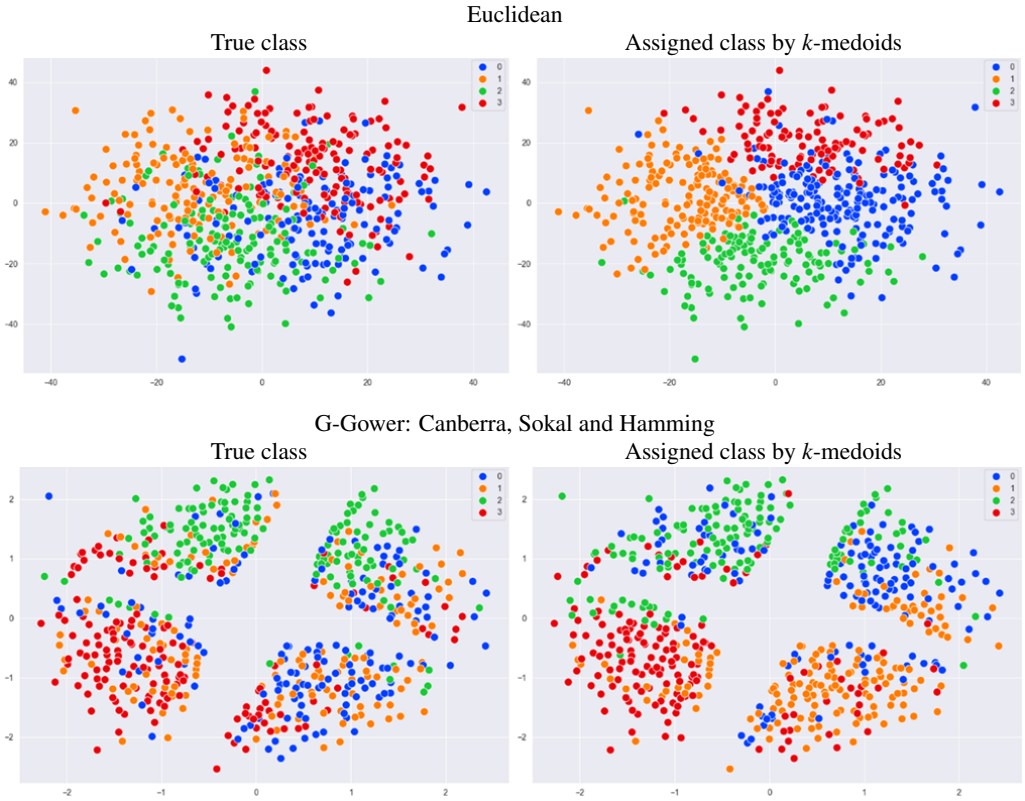

Figure 4. *Clustering visualization. Synthetic dataset 4.*

3.2. Real datasets

3.2.1. Dubai properties dataset

This dataset contains 1905 properties for which 38 characteristics were measured. It is available at `https://www.kaggle.com/datasets/dataregress/dubai-properties-dataset?resource=download`.

The variables considered as predictors are: Latitude and longitude, market price and size (in m^2) as numerical; number of bedrooms (0,1,2,3,4,5) and number of bathrooms (0,1,2,3,4,5,6) as multiclass, and balcony (1=true, 0=false), barbecue are (1=true, 0=false) and private pool (1=true, 0=false) are taken as binary. We decided to consider house quality (1=Low, 0=Medium/High/UltraHigh) as response variable with $k = 2$ classes.

Table 5 contains the classification rate and ARI values for k-medoids algorithm with $k = 2$, where we observe classification rates higher than 85% for some of the robust proposals. In particular, G-Gower with robust Mahalanobis (trimmed, winsorized and MAD), Jaccard and Hamming attain a classification rate of 86.14%, RelMS with robust Mahalanobis (MAD), Jaccard and Hamming reaches the 86.09%, RelMS with robust Mahalanobis (trimmed and winsorized), Jaccard and Hamming reach 85.98% and for

robust Mahalanobis (trimmed, winsorized and MAD), Sokal and Hamming a 85.83% is attained. On the other hand, the classification rate for Euclidean and Gower are 60.58% and 50.97%, respectively.

Table 5. *Classification results for Dubai properties dataset.*

Distance	**Classification rate**	**ARI**
G-Gower: Mahalanobis-Jaccard-Hamming	0.861942	0.505161
G-Gower: Robust Mahalanobis trimmed-Jaccard-Hamming	0.861417	0.503672
G-Gower: Robust Mahalanobis winsorized-Jaccard-Hamming	0.861417	0.503672
G-Gower: Robust Mahalanobis MAD-Jaccard-Hamming	0.861417	0.503672
RelMS: Mahalanobis-Jaccard-Hamming	0.860892	0.502544
G-Gower: Canberra-Jaccard-Hamming	0.860892	0.502365
RelMS: Canberra-Jaccard-Hamming	0.860892	0.502365
RelMS: Robust Mahalanobis MAD-Jaccard-Hamming	0.860892	0.502365
RelMS: Pearson-Jaccard-Hamming	0.860892	0.502365
G-Gower: Pearson-Jaccard-Hamming	0.860367	0.500702
RelMS: Robust Mahalanobis winsorized-Jaccard-Hamming	0.859843	0.499400
RelMS: Robust Mahalanobis trimmed-Jaccard-Hamming	0.859318	0.497922
RelMS: Mahalanobis-Sokal-Hamming	0.858268	0.495329
G-Gower: Robust Mahalanobis MAD-Sokal-Hamming	0.858268	0.495329
RelMS: Pearson-Sokal-Hamming	0.858268	0.495329
RelMS: Robust Mahalanobis trimmed-Sokal-Hamming	0.858268	0.495329
RelMS: Canberra-Sokal-Hamming	0.858268	0.495329
RelMS: Robust Mahalanobis winsorized-Sokal-Hamming	0.858268	0.495329
RelMS: Robust Mahalanobis MAD-Sokal -Hamming	0.858268	0.495329
G-Gower: Robust Mahalanobis winsorized-Sokal-Hamming b	0.858268	0.495329
G-Gower: Mahalanobis-Sokal-Hamming	0.858268	0.495329
G-Gower: Pearson-Sokal-Hamming	0.858268	0.495329
G-Gower: Robust Mahalanobis trimmed-Sokal-Hamming	0.858268	0.495329
RelMS: Euclidean-Jaccard-Hamming	0.819948	0.394762
RelMS: Manhattan-Jaccard-Hamming	0.819948	0.394762
G-Gower: Manhattan-Jaccard-Hamming	0.817848	0.389618
G-Gower: Euclidean-Jaccard-Hamming	0.817848	0.389618
Euclidean	0.605774	0.000912
G-Gower: Canberra-Sokal-Hamming	0.516010	-0.002348
RelMS: Manhattan-Sokal-Hamming	0.510761	-0.000050
RelMS: Euclidean-Sokal-Hamming	0.510761	-0.000050
classical Gower	0.509711	-0.000083
G-Gower: Euclidean-Sokal-Hamming	0.506562	-0.000263
G-Gower: Manhattan-Sokal-Hamming	0.506562	-0.000263

Similar conclusions can be derived regarding ARI. For example, G-Gower with robust Mahalanobis (trimmed, winsorized and MAD), Jaccard and Hamming reach one of the highest ARIs, as well as RelMS with robust Mahalanobis (MAD), Jaccard and Hamming. On the other hand, values around 0 are obtained by classical Gower and Euclidean distance.

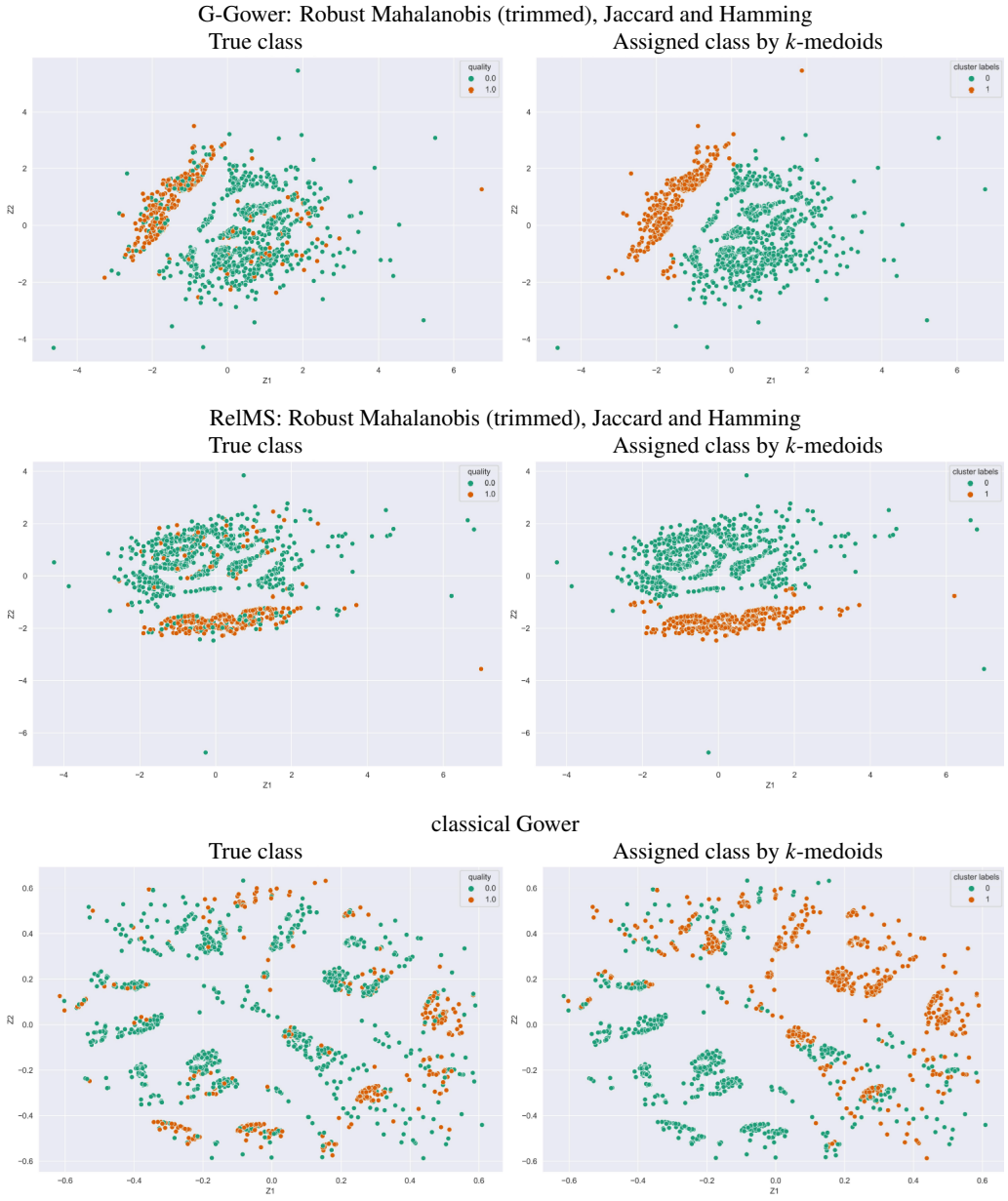

Figure 5. *Clustering visualization. Dubai properties dataset.*

In Figure 5 we use metric MDS maps to visualize the classification found by k-medoids for different metrics ($k = 2$). In particular, we illustrate the results obtained by G-Gower with robust Mahalanobis trimmed, Jaccard and Hamming, RelMS with the same combination of metrics and classical Gower. Units in left panels are colored according to their true class and in right panels, according to their assigned class.

Looking at G-Gower and RelMS panels, we can see that k-medoids is able to identify the class of most of the units, which is coherent with the classification rates obtained, higher than 80%. Focusing on Gower panels, we can see that many units are not well identified, and this is reflected in a classification rate of 51%. Considering that there are only two classes, this means that the classification is practically the same as what would be obtained if the units were classified randomly, following a uniform probability distribution. The ARI value is 0, which is consistent with the low classification rate obtained.

3.2.2. World development indicators dataset

This dataset contains 18 indicators measured 97 countries. It is available at `https://www.kaggle.com/datasets/hn4ever/world-development-indicators-by-countries`. Source: World Bank.

We consider the following predictors. Numerical variables: Access to electricity (% of the population with access to electricity in 2017), life expectancy (number of years a newborn in 2019 would live if the mortality patterns existing at the time of his/her birth remained the same throughout his/her life),insufficient nutrition (% of the population in 2019 whose habitual food consumption is insufficient to provide the levels of dietary energy necessary to maintain a normal active and healthy life), arable land (% of the country's land that is arable in 2014-16), population with less than 3.20$ per day (% of population with less than 3.20$ per day in PPP); Binary variables: Quality of the health system (0=Suitable 1=Inappropriate), Investment in education and health (0=High, 1=Low); multiclass variables: Pollution (0=High, 1=Medium, 2=Low), Insecurity (0=High, 1=Medium, 2=Low). Variable Poverty (0=High, 1=Medium, 2=Low) is taken as response variable with $k = 3$ classes.

Table 6 contains the classification rate and ARI values for k-medoids algorithm with $k = 3$, where we observe classification rates around 65% for some of the robust proposals. In particular, G-Gower with robust Mahalanobis (trimmed and winsorized), Sokal or Jaccard and Hamming attains a 64.95%, RelMS with robust Mahalanobis (trimmed and winsorized), Jaccard and Hamming reaches a 63.92%, as well as G-Gower with robust Mahalanobis (MAD), Sokal or Jaccard and Hamming. On the other hand, the classification rate with Euclidean distance is 46.39% and 60.82% with classical Gower.

Similar results can be observed concerning ARI, where the robust proposals tend to attain ARI values around 0.35-0.37. On the other hand, classical Gower and Euclidean distance obtain values of 0.26 and 0.04, respectively.

Table 6. *Classification results for World development indicators dataset.*

Distance	Classification rate	ARI
G-Gower: Canberra-Jaccard-Hamming	0.659794	0.405188
G-Gower: Manhattan-Jaccard-Hamming	0.649485	0.315937
RelMS: Canberra-Jaccard-Hamming	0.649485	0.397914
G-Gower: Robust Mahalanobis winsorized-Sokal-Hamming	0.649485	0.354318
G-Gower: Robust Mahalanobis trimmed-Sokal-Hamming	0.649485	0.375463
G-Gower: Mahalanobis-Sokal-Hamming	0.649485	0.354318
G-Gower: Pearson-Jaccard-Hamming	0.649485	0.315937
G-Gower: Canberra-Sokal-Hamming	0.649485	0.382536
G-Gower: Robust Mahalanobis winsorized-Jaccard-Hamming	0.649485	0.334477
G-Gower: Robust Mahalanobis trimmed-Jaccard-Hamming	0.649485	0.315937
RelMS: Mahalanobis-Jaccard-Hamming	0.649485	0.334477
G-Gower: Euclidean-Jaccard-Hamming	0.639175	0.353248
RelMS: Pearson-Jaccard-Hamming	0.639175	0.313709
G-Gower: Pearson-Sokal-Hamming	0.639175	0.375701
RelMS: Euclidean-Jaccard-Hamming	0.639175	0.353248
RelMS: Robust Mahalanobis trimmed-Jaccard-Hamming	0.639175	0.334477
RelMS: Robust Mahalanobis winsorized-Jaccard-Hamming	0.639175	0.334477
G-Gower: Robust Mahalanobis MAD-Jaccard-Hamming	0.639175	0.307562
G-Gower: Robust Mahalanobis MAD-Sokal-Hamming	0.639175	0.326646
RelMS: Mahalanobis-Sokal-Hamming	0.628866	0.347343
RelMS: Robust Mahalanobis trimmed-Sokal- Hamming	0.628866	0.325988
RelMS: Robust Mahalanobis winsorized-Sokal-Hamming	0.628866	0.325988
RelMS: Robust Mahalanobis MAD-Jaccard-Hamming	0.628866	0.288271
G-Gower: Mahalanobis-Jaccard-Hamming	0.628866	0.319983
RelMS: Robust Mahalanobis MAD-Sokal-Hamming	0.618557	0.305616
RelMS: Canberra-Sokal-Hamming	0.608247	0.255497
classical Gower	0.608247	0.258923
G-Gower: Euclidean-Sokal-Hamming	0.577320	0.208742
G-Gower: Manhattan-Sokal-Hamming	0.577320	0.208742
RelMS: Manhattan-Sokal-Hamming	0.556701	0.170696
RelMS: Pearson-Sokal-Hamming	0.556701	0.170696
RelMS: Euclidean-Sokal-Hamming	0.556701	0.170696
RelMS: Manhattan-Jaccard-Hamming	0.525773	0.096234
Euclidean	0.463918	0.040735

In Figure 6 we illustrate the results of the k-medoids classification for different metrics ($k = 3$) through metric MDS maps. In particular, the results obtained by RelMS with robust Mahalanobis trimmed, Jaccard and Hamming, and Euclidean distance are shown.

Units in left panels are colored according to their true class and in right panels, according to their assigned class. Looking at RelMS panels, we can see that the classification is rather acceptable, since most of the units in class 0 and half of the units in class 1 are well classified. Note that the classification rate obtained is 64.9%, which is around twice a expected classification rate of 33% that would be attained if the units were classified through a uniform random mechanism. Regarding the Euclidean distance panels, we observe that k-medoids algorithm is not able to identify the units' class, which is coherent with an ARI of 0.04 and a classification rate of 46%.

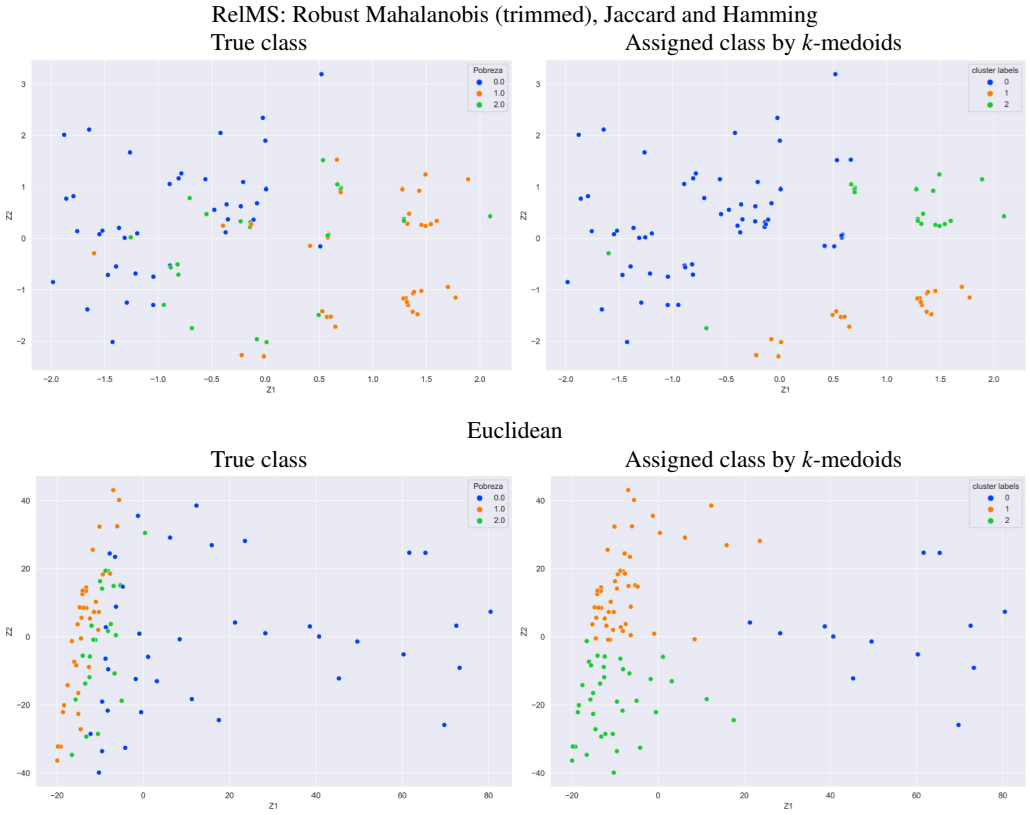

Figure 6. *Clustering visualization. World development indicators dataset.*

3.3. Sensitivity study on the trimming and winsorizing parameter

In this section the sensitivity of parameter α used in trimmed and winsorized versions of the covariance matrix in Mahalanobis distance is studied. Classification rate is used to analyze the performance of the distance for each dataset. Table 7 contains the mean values computed on 100 runs.

Concerning synthetic datasets, we observe that in dataset 1 (10-12% outlier contamination in three numerical variables) G-Gower with robust Mahalanobis (trimmed),

Table 7. *Classification rates for G-Gower and RelMS for several values of trimming and winsoring parameter α.*

Distance	5%	10%	15%	20%	25%
Synthetic dataset 1					
G-Gower: Robust Mahalanobis trimmed-Jaccard-Hamming	0.630	0.628	0.554	0.558	0.556
RelMS: Robust Mahalanobis winsorized-Jaccard-Hamming	0.630	0.628	0.604	0.548	0.544
RelMS: Robust Mahalanobis trimmed-Jaccard-Hamming	0.626	0.598	0.538	0.546	0.546
G-Gower: Robust Mahalanobis winsorized-Jaccard-Hamming	0.604	0.628	0.628	0.556	0.554
RelMS: Robust Mahalanobis trimmed-Sokal-Hamming	0.586	0.584	0.580	0.596	0.594
RelMS: Robust Mahalanobis winsorized-Sokal-Hamming	0.582	0.586	0.580	0.592	0.592
G-Gower: Robust Mahalanobis winsorized-Sokal-Hamming	0.508	0.506	0.508	0.602	0.604
G-Gower: Robust Mahalanobis trimmed-Sokal-Hamming	0.506	0.508	0.504	0.602	0.604
Synthetic dataset 2					
RelMS Robust Mahalanobis trimmed-Jaccard-Hamming	0.491	0.500	0.500	0.502	0.505
RelMS: Robust Mahalanobis trimmed-Sokal-Hamming	0.558	0.558	0.558	0.554	0.555
RelMS: Robust Mahalanobis winsorized-Sokal-Hamming	0.558	0.558	0.558	0.483	0.554
G-Gower: Robust Mahalanobis trimmed-Sokal-Hamming	0.552	0.552	0.552	0.548	0.549
G-Gower: Robust Mahalanobis winsorized-Sokal-Hamming	0.552	0.552	0.552	0.522	0.548
G-Gower -Robust Mahalanobis trimmed-Jaccard-Hamming	0.540	0.540	0.540	0.522	0.525
G-Gower: Robust Mahalanobis winsorized-Jaccard-Hamming	0.540	0.540	0.540	0.449	0.522
RelMS: Robust Mahalanobis trimmed-Jaccard-Hamming	0.480	0.482	0.480	0.554	0.560
RelMS: Robust Mahalanobis winsorized-Jaccard-Hamming	0.477	0.478	0.480	0.442	0.554
Synthetic dataset 3					
G-Gower: Robust Mahalanobis winsorized-Jaccard-Hamming	0.690	0.687	0.822	0.840	0.840
RelMS: Robust Mahalanobis winsorized-Jaccard-Hamming	0.690	0.683	0.792	0.817	0.817
G-Gower: Robust Mahalanobis trimmed-Jaccard-Hamming	0.688	0.833	0.837	0.843	0.685
RelMS: Robust Mahalanobis trimmed-Jaccard-Hamming	0.687	0.815	0.812	0.685	0.687
RelMS: Robust Mahalanobis trimmed-Sokal-Hamming	0.593	0.868	0.878	0.877	0.878
G-Gower: Robust Mahalanobis trimmed-Sokal-Hamming	0.592	0.853	0.885	0.883	0.885
G-Gower: Robust Mahalanobis winsorized-Sokal-Hamming	0.588	0.590	0.865	0.885	0.883
RelMS: Robust Mahalanobis winsorized-Sokal-Hamming	0.585	0.600	0.857	0.878	0.875
Dubai properties dataset					
G-Gower: Robust Mahalanobis trimmed-Jaccard-Hamming	0.861	0.861	0.861	0.862	0.861
G-Gower: Robust Mahalanobis winsorized-Jaccard-Hamming	0.861	0.861	0.861	0.861	0.861
RelMS: Robust Mahalanobis trimmed-Jaccard-Hamming	0.860	0.859	0.860	0.861	0.862
RelMS: Robust Mahalanobis winsorized-Jaccard-Hamming	0.859	0.860	0.860	0.860	0.860
G-Gower: Robust Mahalanobis trimmed-Sokal-Hamming	0.858	0.858	0.858	0.858	0.858
G-Gower: Robust Mahalanobis winsorized-Sokal-Hamming	0.858	0.858	0.858	0.858	0.858
RelMS: Robust Mahalanobis trimmed-Sokal-Hamming	0.858	0.858	0.858	0.858	0.858
RelMS: Robust Mahalanobis winsorized-Sokal-Hamming	0.858	0.858	0.858	0.858	0.858
World development indicators dataset					
G-Gower: Robust Mahalanobis trimmed-Jaccard-Hamming	0.649	0.649	0.649	0.649	0.649
G-Gower: Robust Mahalanobis winsorized-Jaccard-Hamming	0.649	0.649	0.649	0.649	0.649
RelMS: Robust Mahalanobis trimmed-Jaccard-Hamming	0.649	0.649	0.639	0.639	0.639
RelMS: Robust Mahalanobis winsorized-Jaccard-Hamming	0.649	0.649	0.639	0.639	0.649
G-Gower: Robust Mahalanobis trimmed-Sokal-Hamming	0.649	0.649	0.649	0.639	0.639
G-Gower: Robust Mahalanobis winsorized-Sokal-Hamming	0.649	0.649	0.649	0.649	0.649
RelMS: Robust Mahalanobis trimmed-Sokal-Hamming	0.629	0.629	0.629	0.629	0.629
RelMS: Robust Mahalanobis winsorized-Sokal-Hamming	0.629	0.629	0.629	0.629	0.629

Jaccard and Hamming and RelMS with robust Mahalanobis (winsorized), Jaccard and Hamming present the highest classification rate (63.0%), which is attained for $\alpha = 0.05$. The second best rate (62.8%) is attained by the same metrics for $\alpha = 0.10$. In dataset 2 (10% outlier contamination in two numerical variables), the best result (55.8%) is obtained for RelMS with robust Mahalanobis (trimmed and winsorized), Sokal and Hamming for $\alpha = 0.05, 0.10, 0.15$. In dataset 3 (7-8% outlier contamination in three numerical variables), G-Gower with robust Mahalanobis (trimmed and winsorized), Sokal and Hamming reach the best results (88.3%-88.5%) for $\alpha = 0.20, 0.25$.

No general conclusions can be derived from the analysis of the synthetic datasets. Although for datasets 1 and 2 the best results are obtained for α values close to the real proportion of outlying units, this is not the case for dataset 3 where hard trimming/winsorizing is needed. Some explanations may be found in the complexity degree of dataset 3, with unbalanced classes and less redundant information than in datasets 1 and 2.

To better analyze the sensitivity of parameter α classification rates are depicted in Figure 7, where we observe that for dataset 1 the classification rate tends to decrease

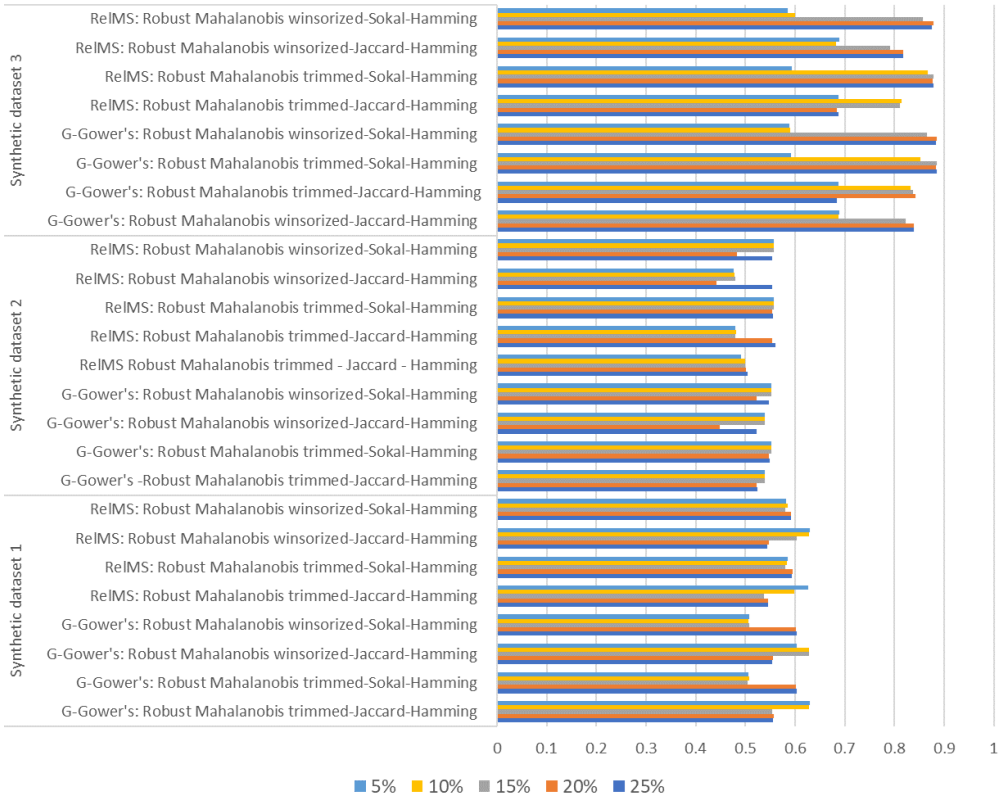

Figure 7. *Classification rates for G-Gower and RelMS for several values of trimming and winsoring parameter α in synthetic datasets.*

with α when Jaccard distance is considered (either in G-Gower or in RelMS), while the opposite happens for Sokal-Michener's. There is not a clear pattern for dataset 2, although the classification rate in G-Gower seems to decrease with α, whereas it stands still or slightly increases when using RelMS. Finally, for dataset 3, the classification rate tends to increase with α in most of the cases considered.

Regarding real datasets, the fluctuation in the classification rate for different values of α is lower than 10^{-2}. In general, for Dubai properties data, the best situation (86.1%) is produced with G-Gower metric with robust Mahalanobis (trimmed and winsorized), Jaccard and Hamming, at any value of α. In the case of World development indicators data, the best results (64.9%) are attained for G-Gower metric with robust Mahalanobis (trimmed and winsorized), Jaccard and Hamming, as well as G-Gower's with robust Mahalanobis (winsorized), Sokal and Hamming at any value of α. The same classification rates are reached for several trimming/winsorizing values of the remaining metrics, except for RelMS with robust Mahalanobis (trimmed and winsorized), Sokal and Hamming, whose classification rate is always equal to 62.9%.

4. Conclusions

In this work, new robust distances for mixed-type data were proposed and studied in the context of clustering. They were obtained as combination of three distances, one for each type of data (numerical, binary and multiclass). As a result, a total of 34 distances were analyzed.

Their performance was evaluated in rather complex synthetic datasets, with underlying correlation structure and outlying units, as well as in two real datasets. Classification rate and adjusted Rand index were used to evaluate the goodness of the clustering obtained with the k-medoids algorithm. Metric multidimensional scaling was used to visualize and illustrate the difference between the true and assigned class of the units. In all scenarios with underlying correlation structure and outlying units, new robust proposals outperformed the classical Gower distance. In the absence of correlation or outlying units, Euclidean distance achieved the best results. However, this is not the usual context in real-world applications, where outliers are highly probable and redundant information is often present in multivariate data. In addition, a sensitivity analysis on the parameters involved in the robust estimation of the new proposals was provided. Some of the robust proposals became computationally unfeasible for sample sizes larger than 30000, with an i7-1365U 1.80 GHz processor, 32.0 GB RAM, where no parallelization was used. The study of their feasibility in larger sample sizes is left for further research.

5. Software availabily

All the distances presented in this paper are implemented in a Python package, called robust_mixed_dist, hosted in `https://pypi.org/project/robust_mixed_dist/`. The package is described in Appendix A and a tutorial is available at `https://fabios`

cielzoortiz.github.io/robust_mixed_dist-docu/intro.html. The robust_mixed_dist package relies on Scipy for an efficient computation of distances. Thus, all distance functions available therein can be easily included in our package. However, the handling of missing data is not considered in Scipy and we leave it for further research.

Funding

The authors acknowledge the support of grants PID2021-123592OB-I00 and TED2021-129316B-I00, funded by MICIU/AEI/10.13039/501100011033, "ERDF A way of making Europe" and "European Union NextGenerationEU/PRTR".

Acknowledgments

The authors thank anonymous reviewers and AE for their comments, leading to an improved version of the manuscript. Special thanks to Prof. Pierre Legendre for his constructive and detailed comments, suggestions, and references.

References

Albarrán, I., Alonso, P. and Grané, A. (2015). Profile identification via weighted related metric scaling: an application to dependent spanish children. *Journal of the Royal Statistical Society - Series A. Statistics in Society*, 178:1–26.

Boj, E. and Grané, A. (2024). The robustification of distance-based linear models: some proposals. *Socio-Economic Planning Sciences*, 95:11992.

Borg, I. and Groenen, P. (2005). *Modern multidimensional scaling: theory and applications*. Springer, New York.

Cuadras, C.M. and Fortiana, J. (1995). A continuous metric scaling solution for a random variable. *Journal of Multivariate Analysis*, 52:1–14.

Cuadras, C.M. and Fortiana, J. (1998). Visualizing categorical data with related metric scaling. In: *Visualization of Categorical Data*. Ed. by J. Blasius and M. Greenacre. London: Academic Press, pages 365–376.

Devlin, S.J., Gnanadesikan, R. and Kettenring, J.R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62(3):531–545.

Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*. John Wiley and Sons, London.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysisilarity coefficients. *Biometrika*, 53:325–338.

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrika*, 27:857–874.

Gower, J.C. and Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3:5–48.

Grané, A., Manzi, G. and Salini, S. (2022). Dynamic mixed data analysis and visualization. *Entropy*, 24:1399.

Grané, A. and Romera, R. (2018). On visualizing mixed-type data: A joint metric approach to profile construction and outlier detection. *Sociological Methods and Research*, 47(2):207–39.

Grané, A., Salini, S. and Verdolini, E. (2021). Robust multivariate analysis for mixed-type data: Novel algorithm and its practical application in socio-economic research. *Socio-Economic Planning Sciences*, 73:100907.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société vaudoise des sciences naturelles*, 37:547–579.

Kaufman, L. and Rousseeuw, P. (1990). *Finding groups in data: an introduction to cluster analysis*. Wiley, Boca Raton.

Legendre, P. and De Cáceres, M. (2013). Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecology Letters*, 16:951–963.

Nguyen X.V., Epps, J. and Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, 1073–108.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 53:846–850.

Sokal, R.R. and Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

A. The Python package

The robust_mixed_dist package is a new Python tool for computing classical statistical distances between units. The distance functions implemented are: Euclidean (ℓ^2 distance), Minkowski (family of ℓ^p distances), Canberra, Pearson (standardized Euclidean), Mahalanobis, robust Mahalanobis, Gower, generalized Gower (G-Gower) and related metric scaling (RelMS). A total of 41 statistical distances can be calculated, including those proposed in this paper. Additionally, since robust_mixed_dist relies on Scipy Python library, all distances included there can be easily included. Forn example, this is the case for Chi-square distance, Chebyschev distance, Jensen-Shannon distance, among others, as well as many similarity coefficients available at Scipy.

The package is hosted in PyPI (Python Packages Index), the official repository of Python packages. More information about robust_mixed_dist can be found in `https://pypi.org/project/robust_mixed_dist/`.

In what follows we provide a small demonstration of how to use robust_mixed_dist. For more details, a more extensive tutorial is available at `https://fabioscielzoortiz.github.io/robust_mixed_dist-docu/intro.html`.

Example of use

Installing:

```
pip install robust-mixed-dist
```

Loading modules:

```
from robust_mixed_dist.mixed import GGowerDistMatrix
from robust_mixed_dist.mixed import RelMSDistMatrix
from robust_mixed_dist.quantitative import robust_maha_dist_matrix, S_robust
```

Computing some distances for a real Madrid houses dataset, which can be found at https://www.kaggle.com/datasets/mirbektoktogaraev/madrid-real-estate-market. In this brief tutorial only the following variables of that dataset were considered:

- Quantitative: sq mt built, number of rooms, number of bathrooms, number of floors, buy price.
- Binary: is renewal needed, has lift, is exterior, has parking.
- multiclass: energy certificate, house type.

Robust Mahalanobis (MAD):

```
S_robust_ = S_robust(X=madrid_houses_df, method='MAD',
                     epsilon=0.05, n_iters=20)
robust_maha_dist_matrix(madrid_houses_df, S_robust=S_robust_)
```

```
array([[ 0.        ,  6.47092419,  7.01983235, ...,  4.96377088,
         5.69177645,  3.68021705],
       [ 6.47092419,  0.        ,  3.03471006, ..., 10.43356417,
        10.12781147,  5.95613137],
       [ 7.01983235,  3.03471006,  0.        , ..., 11.35024985,
        10.9171085 ,  6.21243845],
       ...,
       [ 4.96377088, 10.43356417, 11.35024985, ...,  0.        ,
         3.65216542,  7.11373136],
       [ 5.69177645, 10.12781147, 10.9171085 , ...,  3.65216542,
         0.        ,  7.86440327],
       [ 3.68021705,  5.95613137,  6.21243845, ...,  7.11373136,
         7.86440327,  0.        ]])
```

G-Gower: Robust Mahalanobis (trimmed), jaccard, Hamming (matching):

```
GG_init = GGowerDistMatrix(p1=5, p2=4, p3=2,
           d1='robust_mahalanobis', d2='jaccard', d3='hamming',
           method='trimmed', alpha=0.05, epsilon=0.05, n_iters=20,
           fast_VG=False)
D_GG = GG_init.compute(X=madrid_houses_df)
D_GG
```

```
array([[0.       , 2.21885363, 1.93318704, ..., 1.93891555, 3.12986955,
        2.26834878],
       [2.21885363, 0.       , 1.22309875, ..., 2.38689136, 2.63841547,
        2.01262089],
       [1.93318704, 1.22309875, 0.       , ..., 2.3585878 , 2.50589448,
        1.63422771],
       ...,
       [1.93891555, 2.38689136, 2.3585878 , ..., 0.       , 2.89514966,
        1.7665964 ],
       [3.12986955, 2.63841547, 2.50589448, ..., 2.89514966, 0.       ,
        3.02408907],
       [2.26834878, 2.01262089, 1.63422771, ..., 1.7665964 , 3.02408907,
        0.       ]])
```

RelMS: Robust Mahalanobis (winsorized), Jaccard, Hamming (matching):

```
RelMS_init = RelMSDistMatrix(p1=5, p2=4, p3=2,
                d1='robust_mahalanobis', d2='jaccard', d3='hamming',
                method='winsorized', epsilon=0.05, alpha=0.05,
                n_iters=20)
D_RelMS = RelMS_init.compute(X=madrid_houses_df.head(1000),
                             Gs_PSD_trans=True)
D_RelMS
```

```
array([[ 0.       , 10.29131982, 10.20148541, ..., 10.25180831,
         10.1963865 , 10.23458336],
       [10.29131994,  0.       , 10.13419898, ..., 10.10288394,
         10.08333674, 10.0731428 ],
       [10.20148539, 10.134199  ,  0.       , ..., 10.14619431,
         10.03394396, 10.11537897],
       ...,
       [10.25180831, 10.10288394, 10.14619431, ...,  0.       ,
         10.05982432, 10.00909222],
       [10.1963865 , 10.08333674, 10.03394396, ..., 10.05982432,
          0.       , 10.03008924],
       [10.23458336, 10.0731428 , 10.11537897, ..., 10.00909221,
         10.03008923,  0.       ]])
```

B. Computational cost

In this section a computational experiment is derived to illustrate the computational cost of G-Gower and RelMS metrics.

For such purpose eight synthetic datasets were generated using `make_blobs` function from `scikit-learn` Python library. The datasets have increasing sample size, going from $n = 3{,}000$ to $n = 70{,}000$, and same characteristics like $p_1 = 4$ numerical, $p_2 = 2$ binary and $p_3 = 2$ multiclass variables and number of true classes $k = 3$.

For each dataset the computation time for G-Gower (robust Mahalanobis trimmed-Jaccard-Hamming), RelMS (robust Mahalanobis trimmed-Jaccard-Hamming) and robust Mahalanobis (trimmed) distances was collected. The experiment was carried out with a DESKTOP-I1Q2RCC device, 13th Gen Intel(R) Core(TM) i7-1365U 1.80 GHz processor, with Installed RAM of 32.0 GB (31.6 GB usable) and 64-bit operating system, x64-based processor. Results are shown in Figure 8.

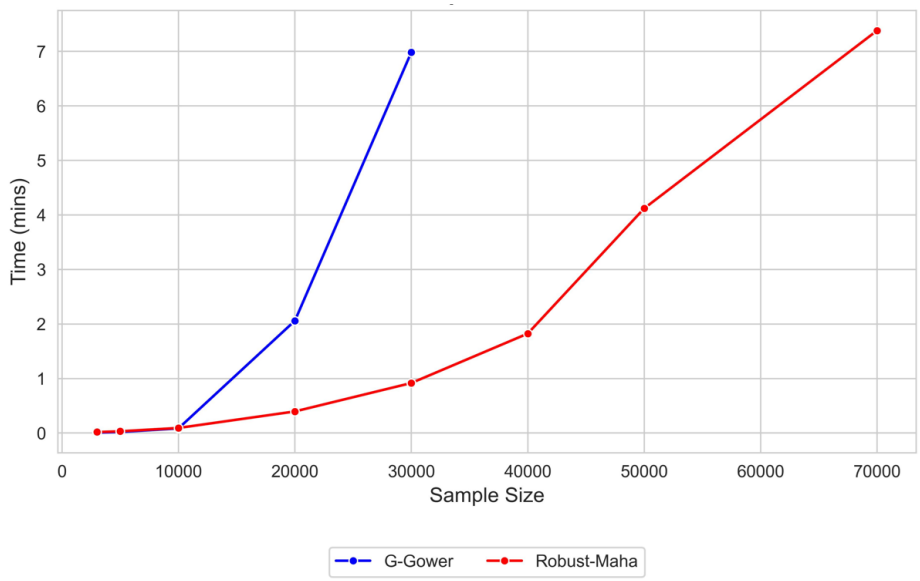

Figure 8. *Computational time (in minutes) for G-Gower and robust Mahalanobis.*

The computationally cheapest distance is robust Mahalanobis, since it takes less than one minute for $n = 30,000$ and below, between one and four minutes for sizes between $n = 30,000$ and $n = 50,000$, and for the largest size ($n = 70,000$) it takes seven minutes, approx. Its calculation is feasible for all sample sizes tested, and reasonably practical for all of them as well. G-Gower takes few seconds for sample sizes up to $n = 10,000$, no more than two minutes for sample sizes between $n = 10,000$ and $n = 20,000$, between 2 and 7 minutes for sizes between $n = 20,000$ and $n = 30,000$, and becomes unfeasible for larger sample sizes. Computational time for RelMS goes from 2.37 minutes for $n = 3,000$ to 534.07 minutes for $n = 20,000$, this is the reason why they are not included in Figure 8. Sample sizes $n \geq 30,000$ are unfeasible for RelMS.

SORT 49 (2) July-December 2025, 245-264 DOI: 10.57645/20.8080.02.29

Bayesian estimation for conditional probabilities associated to directed acyclic graphs: study of hospitalization of severe influenza cases

Lesly Acosta[1,*] and Carmen Armero[2]

Abstract

This paper presents a Bayesian framework to estimate joint, conditional, and marginal probabilities in directed acyclic graphs to study the progression of hospitalized patients with confirmed severe influenza. Using data from the PIDIRAC retrospective cohort in Catalonia, we model patient pathways from admission to discharge, death, or transfer. Transition probabilities are estimated using a Bayesian Dirichlet-multinomial approach, while posterior distributions for absorbing states or inverse probabilities are assessed via simulation. Bayesian methodology quantifies uncertainty through posterior distributions, offering insights into disease progression and in improving hospital planning. These findings support more effective patient management and informed decision making during seasonal influenza outbreaks.

MSC: *62F15, 62P10.*

Keywords: *Confirmed influenza hospitalization, Directed acyclic graphs (DAGs), Dirichlet-multinomial Bayesian inferential process, Healthcare decision-making, Transition probabilities.*

1. Introduction

According to the World Health Organization (WHO): "Seasonal influenza (the flu) is an acute respiratory infection caused by influenza viruses common in all parts of the world". Following official estimates, about 1 billion cases of seasonal influenza occur worldwide each year. This includes between 3 and 5 million cases of severe illness and between 290 and 650 thousand respiratory deaths caused by the disease (Pietrasik, 2023). Influenza rates are very high in children but its mortality records are shocking in elderly

* Corresponding author: Lesly Acosta; lesly.acosta@upc.edu

[1] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona-TECH.

[2] Department of Statistics and Operations Research, Universitat de València, Burjassot, Spain.

Received: April 2025
Accepted: October 2025

populations as well as in people affected by chronic diseases (Paget et al., 2022). It is a worldwide cause of hospital admissions and mortality in these latter groups (Macias et al., 2021).

Influenza prevention and control remains a serious public health challenge, despite the availability of vaccines and antiviral treatments (Carrillo-Santisteve et al., 2012). The European Centre for Disease Prevention and Control (ECDC) is an agency of the European Union (EU) that collects epidemiological and virological data from member countries of the European Economic Area (EEA). Surveillance data come from the sentinel influenza surveillance systems of each associated country, which may cover substantial parts of the population or even have a universal surveillance system (Snacken and Brown, 2015).

Although most people affected by the seasonal influenza recover within one or two weeks without medical attention, it can cause serious illnesses and mortality, especially among population at higher risk. Severe influenza complications can result in hospitalization, possibly with admission to the ICU or even death (Acosta et al., 2021; Soldevila et al., 2021). According to ECDC, around 10 to 30% of Europe's population is infected annually with influenza, causing hundreds of thousand hospitalizations. A systematic review of the clinical burden of influenza disease in older people was done by Langer et al. (2023) with data from January 2012 to February 2022.

There are almost no studies that quantitatively analyse the different health conditions that hospitalised patients with severe influenza can experience from admission to discharge. Knowledge of these pathways would be a valuable tool for improving hospital resource planning and organization of the seasonal period of influenza, the winter. In Europe, influenza generally causes annual epidemics that affect up to 20% of the population.

Graph theory is a very theoretical mathematical subject with an enormous power to visualize the basic functioning of scenarios that operate in environments with many sources of uncertainty. Our approach to graphs is essentially graphical and structural. In particular, the evolution of a hospitalized patient from admission to discharge can be represented graphically by means of a probabilistic directed acyclic graph (DAG) (Cowell et al., 1999; Barber, 2012) with nodes defined by random events associated to the different health conditions of the inpatients and arrows connecting two consecutive nodes without any possibility of return. Transition probabilities between nodes are conditional probabilities that provide valuable clinical information on the state of health in which individuals move from their current status. They are the basis for assessing the uncertainty associated with the different trajectories of the study, the final (absorbing) states, and inverse probabilities that inform previous events.

This paper presents a general inferential procedure for estimating joint, conditional, and marginal probabilities in probabilistic DAGs associated to random events, which we apply to assess the different pathways that a patient with severe influenza may follow from their admission to the hospital to their discharge, as fully cured, dead or sent to a long-stay facility. These latter type of institutions usually welcome patients with chronic

diseases and comorbidities who have little hope of cure, but require continued clinical care.

All the inferential processes in this work are framed within the Bayesian inferential methodology. This will allow to directly quantify the uncertainty associated with the relevant outcomes through probability distributions. In particular, posterior distribution associated to transition probabilities or absorbing states will allow us to better understand the hospital evolution of patients with severe influenza. Overall, they may be a useful tool in the effective management of patients hospitalised with influenza during peaks of influenza epidemic activity.

This paper is organized as follows. Section 2 presents a description of the motivated problem, named the PIDIRAC cohort study, and data as well as the DAG that represents the evolution of a severe influenza patient in hospital. Section 3 introduces the general Bayesian modeling for assessing conditional probabilities for adjacent and non adjacent states of a DAG. Section 4 applies the general approach in Section 3 to the data of the PIDIRAC study. Section 5 provides a more general framework for modelling conditional probabilities through covariates and illustrates its application in a particular probability distribution from the PIDIRAC project. Finally, Section 6 contains some general conclusions and comments.

2. PIDIRAC retrospective cohort study

This study focuses on hospitalized severe influenza patients. Data for the analysis were collected from a retrospective cohort study of hospitalized, laboratory confirmed, influenza (SHLCI) patients registered from 1 October 2017 to 22 May 2018 by the 14 hospitals included in the Primary Care Influenza Surveillance System of Catalonia (PIDIRAC). Catalonia is an autonomous community of Spain located in the northeast of the Iberian Peninsula with a population of around 8 million inhabitants. The median and interquartile range age of the hospitalized patients was 72 and 59−83 years respectively, with 563 (43%) of all being female.

All severe influenza patients who came to the hospital were initially attended by a physician who, depending on the patient's state of health, recommended admission to an intensive care unit (ICU, denoted by I when treated as a variable) or to a specific hospital ward ($W1$). Some of the patients who were initially directed to $W1$ were later transferred to the ICU, derived to a long-term care facility (L), died (D) or were cured and discharged from the hospital and sent home (H). Patients in ICU can die or, if they improved, be sent to a second type of ward ($W2$) of the hospital, from where they can move to H, L or D.

Consider a graph $G = (S, \mathcal{C})$ with nodes $S = \{A, I, W1, W2, D, H, L\}$ the different states or services in the hospital and ordered arcs

$$\mathcal{C} = \{AI, AW1, W1I, IW2, W1D, W1H, W1L, W2D, W2H, W2L\}$$

connecting the different adjacent nodes. See Figure 1 for a DAG that represents the evolution of hospitalized patients with severe influenza through nodes describing their

different health states and/or hospital services used and the different directed arcs connecting neighbouring nodes. The nature of the different states is different: The admission of a patient to the hospital is the initial state A, states I, $W1$, and $W2$ are transient, and states D, H, and L absorbing.

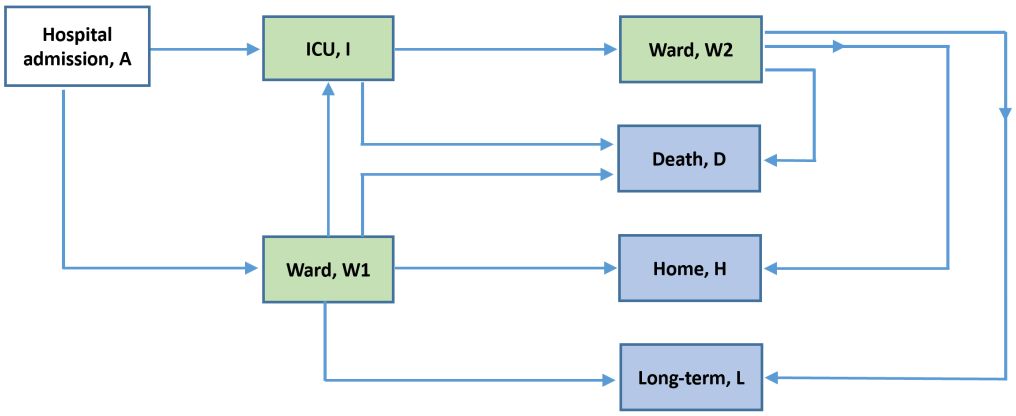

Figure 1. *Directed acyclic graph for a description of the progress of patients with severe influenza in hospital. Transient states in green and absorbing states in blue.*

Over the study period, the number of patients with a diagnosis of severe influenza admitted to the hospital was 1306. Of these patients, 1208 were initially referred to the ward and 98 were sent to the ICU. A total of 82 patients in $W1$ were subsequently transferred to the ICU. Of the total of 1126 patients on the ward not sent to the ICU, 946 were discharged and sent home, 55 were sent to L, and 125 died. A total of 35 patients died during their stay in ICU; and the rest, 145, were sent to $W2$ from where 118 were discharged and sent home, 12 were sent to a long-stay facility and 15 died.

3. Bayesian modeling of conditional probabilities

The probabilistic approach to our model is concentrated on conditional probabilities that assess the uncertainty associated to the different paths between the states of the process: probabilities of visiting a certain state from another given state, directly without intermediate states or not; probabilities of ending in each of the absorbing states; and even, inverse probabilities that focus on assessing the uncertainty associated with a previous state knowing that the process has ended up visiting a certain state subsequently.

3.1. Dirichlet-multinomial learning process for assessing direct conditional probabilities

Assume a graph with a finite set S of states and consider (without loss of generality) that from the state $i \in S$ it is possible to directly visit the states $\{1,\ldots,J\} \in S$ with probabi-

lities $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iJ})^\intercal$, where θ_{ij}, $i \neq j$, represents the probability that the individual's departure from state i will be to visit state j directly, with $\theta_{iJ} = 1 - \sum_{j=1}^{J-1} \theta_{ij}$.

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^\intercal$ be the random multinomial vector whose component j, Y_{ij}, describes the number of individuals that move directly from state i to state j for the n_i individuals that are in the state i and eventually leave it, where $Y_{iJ} = n_i - \sum_{j=1}^{J-1} Y_{ij}$. The parametric vector associated to this multinomial distribution is $\boldsymbol{\theta}_i$. As an illustrative example, in the PIDIRAC study, a total of 145 patients were recorded in state $W2$, who subsequently transitioned to states H, L, or D. In this case, the multinomial outcome vector of interest, $\boldsymbol{Y}_{W2}$, is three-dimensional, with components Y_{W2H}, Y_{W2L}, and Y_{W2D} representing the numbers of individuals in $W2$ who moved to H, L, and D, respectively. The corresponding parameter vector for inference, $\boldsymbol{\theta}_{W2}$, consists of the transition probabilities from $W2$ to H, L, and D, denoted by θ_{W2H}, θ_{W2L}, and θ_{W2D}, respectively.

The basic Bayesian inferential process for multinomial probabilities $\boldsymbol{\theta}_i$ account for the Dirichlet family of distributions as the conjugate family for the sampling multinomial distribution, $\boldsymbol{Y}_i \mid \boldsymbol{\theta}_i \sim \text{Mn}(n_i, \boldsymbol{\theta}_i)$. This implies that if a Dirichlet distribution $\text{Di}(\boldsymbol{\alpha}_i)$ with parameters $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iJ})^\intercal$ is chosen as a prior distribution $\pi(\boldsymbol{\theta}_i)$ for $\boldsymbol{\theta}_i$, the corresponding posterior distribution $\pi(\boldsymbol{\theta}_i \mid \mathcal{D})$, where $\mathcal{D}$ stands for the data, will also be a Dirichlet distribution, $\text{Di}(\boldsymbol{y}_i + \boldsymbol{\alpha}_i)$, where $\boldsymbol{y}_i$ is the vector of the observations from $\boldsymbol{Y}_i$. Just as the binomial distribution is the univariate version of the multinomial distribution, the Beta distribution is the univariate version of the Dirichlet distribution. In this regard, the posterior marginal distribution associated to probability θ_{ij}, $\pi(\theta_{ij} \mid \mathcal{D})$, is a Beta distribution with parameters $y_{ij} + \alpha_{ij}$ and $\alpha_i^+ + n_i - (y_{ij} + \alpha_{ij})$, where $\alpha_i^+ = \sum_{j=1}^{J} \alpha_{ij}$ (Congdon, 2005; Armero et al., 2021).

There are many studies where prior information cannot be provided or is available but it is not intended to be used in the inferential process in order to make the analysis more "objective" (Alvares et al., 2018). In these scenarios, the Bayesian inferential protocol establishes as necessary the specification of a prior distribution that serves as a starting point for the inference and that disturbs and distorts as little as possible the information provided by the data. The choice of such scarce informative distribution is not unanimous in the scientific literature and has generated a lot of controversy. We will not go into this issue and will use the Perk's prior Dirichlet distribution (Perks, 1947; Berger, Bernardo and Sun, 2015), whose parameters share the unit of probability equally and is the most widely used Dirichlet distributions for this type of problem. The corresponding posterior distribution is $\text{Di}(\boldsymbol{y}_i + \mathbf{1}/J)$, where $\mathbf{1}$ is now a vector of ones of dimension J.

3.2. Bayes inference for non-direct conditional probabilities

A little more notation should be added to the study to analyse conditional probabilities associated with non-contiguous states. We define $\theta_{ik_1 \ldots k_K j}$ as the probability associated with the path that begins in state i, then visits state k_1, immediately after state k_2, and so on up to k_K, and finally ends in j, where each of the visited states is temporally following its immediate preceding. In the event the state i is temporally earlier than state j, $\theta_{i\bullet j}$

shall stand for the probability of visiting j from i via any path $\mathcal{P}(i,j)$ that connects both states,

$$\theta_{i\bullet j} = \sum_{\mathcal{P}(i,j)} \theta_{\mathcal{P}(i,j)}. \tag{1}$$

Analogously, if state i is temporally later than state j, $\theta_{i\bullet j}$ will represent the probability of having departed from state j knowing that the process has visited the posterior state i. This probability is calculated from Bayes's theorem as follows

$$\theta_{i\bullet j} = \frac{\theta_{j\bullet i}\,\theta_j}{\theta_i},$$

where now θ_i (θ_j) indicates the probability of visiting state i (j) from the initial state of the process.

The conjugate Dirichlet-multinomial learning process is suitable for the computation of posterior distributions associated to jumping probabilities between a state and its immediate next states, but not for posterior distributions associated to probabilities between non-neighbouring states or inverse probabilities. In this sense, we will assume a Markovian structure for the transition probabilities as follows:

$$\theta_{ik_1\ldots k_K j} = \theta_{ik_1}\,\theta_{k_1k_2}\cdot\ldots\cdot\theta_{k_K j}. \tag{2}$$

This condition allows the simulation of the posterior distribution of the probability associated to this particular trajectory $\pi(\theta_{ik_1\ldots k_K j} \mid \mathcal{D})$. In fact, a simulated random sample $\{\theta^{(m)}_{ik_1\ldots k_K j}, m = 1,\ldots,M\}$ from this posterior distribution is constructed as follows

$$\theta^{(m)}_{ik_1\ldots k_K j} = \theta^{(m)}_{ik_1}\,\theta^{(m)}_{k_1k_2}\cdot\ldots\cdot\theta^{(m)}_{k_K j},$$

where each simulated value $\theta^{(m)}_{k_ik_j}$ is generated from the Beta marginal posterior distribution, $\pi(\theta_{k_ik_j} \mid \mathcal{D})$, of the transition probability from state k_i to its immediate next state k_j.

In this way, we can compute a random sample from $\pi(\theta_{i\bullet j} \mid \mathcal{D})$ taking into account that

$$\pi(\theta_{i\bullet j} \mid \mathcal{D}) = \pi(\textstyle\sum_{\mathcal{P}(i,j)} \theta_{\mathcal{P}(i,j)} \mid \mathcal{D}). \tag{3}$$

In the case that i is subsequent to j, we can generate a random sample from $\pi(\theta_{i\bullet j} \mid \mathcal{D})$ bearing in mind that

$$\pi(\theta_{i\bullet j} \mid \mathcal{D}) = \pi\Big(\frac{\theta_{j\bullet i}\,\theta_j}{\theta_i} \mid \mathcal{D}\Big). \tag{4}$$

The software R, version 4.4.2 (R Core Team, 2025), has been used to implement the computation of all the transition probabilities results reported in this work. The code and the suplementary material can be found in a GitHub repository (https://github.com/LAcosta15/CALA_SupplementaryMaterial).

4. Hospitalization of severe influenza cases

4.1. Probabilities associated with direct transitions between states

We begin the inferential process of the PIDIRAC study by estimating the probability distribution associated with visiting each of the states that can be accessed from an immediately preceding state. In our research these would be the random vector $\boldsymbol{\theta}_A = (\theta_{AW1}, \theta_{AI})^\intercal$ that assesses the probability that a patient admitted to the hospital will be referred to ward $W1$ or ICU, $\boldsymbol{\theta}_I = (\theta_{IW2}, \theta_{ID})^\intercal$ that accounts for the probability that a patient in the ICU moves to ward $W2$ or dies, $\boldsymbol{\theta}_{W1} = (\theta_{W1I}, \theta_{W1D}, \theta_{W1H}, \theta_{W1L})^\intercal$ as the vector that indicates the possible visits to I, D, H and L from $W1$, and the vector $\boldsymbol{\theta}_{W2} = (\theta_{W2D}, \theta_{W2H}, \theta_{W2L})^\intercal$ formed by the probability that a patient ends up in D, H or L since his second stay on the ward, $W2$. In all these cases, we will use the Perk prior distribution introduced above as the non-informative prior distribution of all the inferential processes in our study.

From admission to the ICU or ward $W1$

Data for learning about the probability $\boldsymbol{\theta}_A = (\theta_{AW1}, \theta_{AI})^\intercal$ that a patient admitted to the hospital will be referred to ward $W1$ or ICU, respectively, refer to the number of patients admitted to the hospital, 1306, and how many of them were sent to the ward $W1$ or ICU, 1208 and 98, respectively. The posterior distribution of $\boldsymbol{\theta}_A$ is the Dirichlet distribution

$$\pi(\boldsymbol{\theta}_A \mid \mathcal{D}) = \text{Di}(1208.5, 98.5),$$

with posterior expectations 0.925 and 0.075 for θ_{AW1} and θ_{AI}, respectively, thus showing that about the 92.5% of hospital patients are transferred directly to $W1$, and the rest, around 7.5%, are sent directly to the ICU. Posterior 95% credible intervals for these probabilities are $(0.910, 0.938)$ and $(0.062, 0.090)$, respectively. These are very narrow intervals indicating very little uncertainty about both probabilities. It is interesting to note that these credible intervals provide a direct measure of the uncertainty of θ_{AW1} and θ_{AI} that is not possible to consider in the frequentist framework. Figure 2 shows the marginal beta posterior distribution associated for each of the two probabilities, $\pi(\theta_{AI} \mid \mathcal{D}) = \text{Be}(98.5, 1208.5)$ and $\pi(\theta_{AW1} \mid \mathcal{D}) = \text{Be}(1208.5, 98.5)$.

From ward $W1$ to ICU, death, home or a long-stay facility

Of the 1208 patients who were initially sent to ward $W1$, 82 were transferred to ICU, 946 discharged home, 55 derived to a long-term care facility L, and 125 died. This information generates the following posterior distribution for $\boldsymbol{\theta}_{W1} = (\theta_{W1I}, \theta_{W1D}, \theta_{W1H}, \theta_{W1L})^\intercal$, the vector of the probability associated to a possible visit from $W1$ to I, D, H and L, respectively

$$\pi(\boldsymbol{\theta}_{W1} \mid \mathcal{D}) = \text{Di}(82.25, 125.25, 946.25, 55.25),$$

with posterior expectation and 95% credible interval for the posterior distribution associated with each one of the probabilities in $\boldsymbol{\theta}_{W1}$ in Table 1.

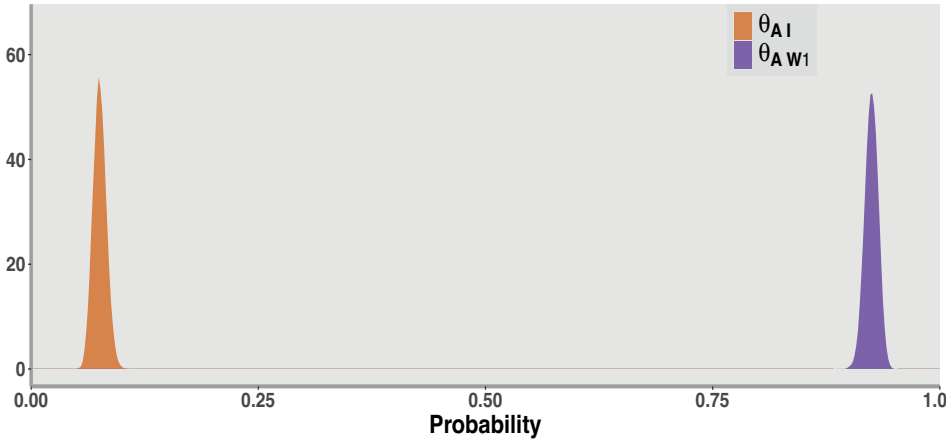

Figure 2. *Posterior marginal distribution for the probability of a patient admitted to the hospital being sent to ICU, θ_{AI}, or ward $W1$, θ_{AW1}.*

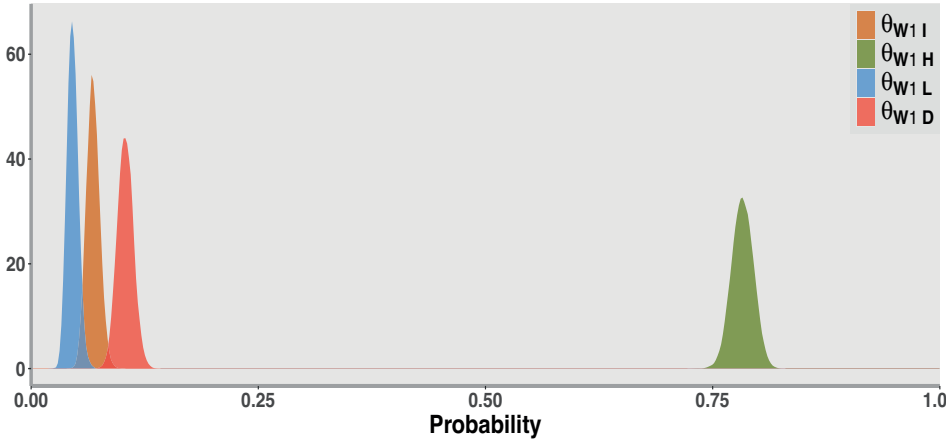

Figure 3. *Posterior conditional distribution for the probability of a patient in $W1$ is sent to a long stay facility, θ_{W1L}, to ICU, θ_{W1I}, to death, θ_{W1D}, or home, θ_{W1H}.*

Table 1. *Posterior distribution, mean, and 95% credible interval for the probabilities in $\boldsymbol{\theta}_{W1}$.*

Probability	Marginal	Mean	CI 95%
θ_{W1I}	Be(82.25, 1126.75)	0.068	(0.055, 0.083)
θ_{W1D}	Be(125.25, 1083.75)	0.103	(0.087, 0.121)
θ_{W1H}	Be(946.25, 262.75)	0.783	(0.759, 0.806)
θ_{W1L}	Be(55.25, 1153.75)	0.046	(0.035, 0.058)

Figure 3 shows the posterior distribution of each of the probabilities associated with $\boldsymbol{\theta}_{W1}$. It is interesting to note the homogeneity of the distributions associated with moving to states I, D and L and the difference in magnitude and variability of that associated with H. Approximately 10% of the patients in the ward $W1$ die, 5% are sent to the ICU, and around 4% are transferred to a long-term institution, probably with very little chance of cure. On the other hand, about 78% of patients admitted to the ward are discharged, although the uncertainty associated with this probability is greater than the previous three.

From ICU to death or to a second hospital ward, $W2$

Among the 180 patients who went through the ICU 35 died, and the rest, 145, were referred to $W2$. As a result, the posterior distribution for $\boldsymbol{\theta}_I = (\theta_{IW2}, \theta_{ID})^\intercal$ that accounts for the probability that a patient in the ICU moves to ward $W2$ or dies is as follows,

$$\pi(\boldsymbol{\theta}_I \mid \mathcal{D}) = \text{Di}(145.50, 35.50),$$

with posterior expectation 0.804 and 95% credible interval (0.743, 0.858) for θ_{IW2}, and 0.196 and (0.142, 0.257) for θ_{ID}, respectively. Figure shows both posterior distributions, very different in their location but similar in shape and variability.

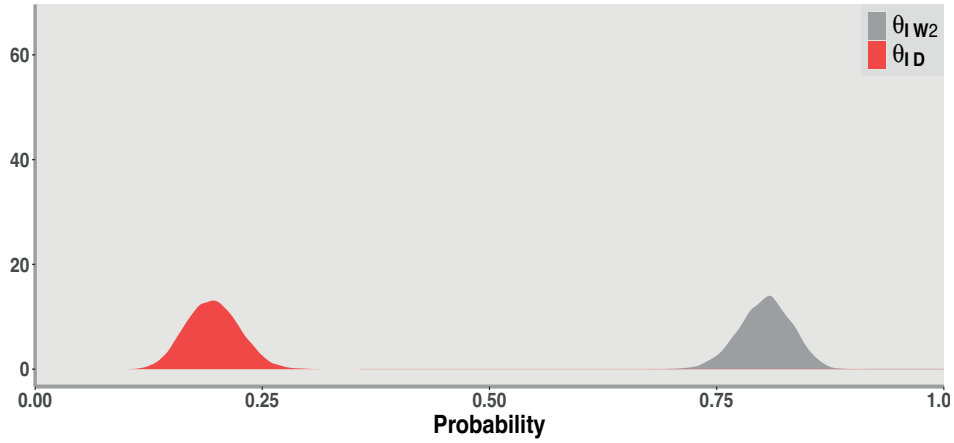

Figure 4. *Posterior distribution of the probability that a patient leaving ICU is transferred to ward,* θ_{IW2} *or dies,* θ_{ID}.

From ward W2 to death, home or a long-term care facility

A total of 145 patients passed through ward $W2$. Of these, 15 died, 118 were discharged and sent home, and the remaining 15 were sent to a long-term care facility. With this information, the posterior distribution associated to $\boldsymbol{\theta}_{W2} = (\theta_{W2D}, \theta_{W2H}, \theta_{W2L})^\intercal$ is:

$$\pi(\boldsymbol{\theta}_{W2} \mid \mathcal{D}) = \text{Di}(145.33, 118.33, 12.33),$$

with posterior mean and 95% credible interval for the posterior distribution associated to each one of the probabilities in $\boldsymbol{\theta}_{W2}$ given in Table 2.

Table 2. *Posterior marginal distribution, mean and 95% credible interval for the probabilities in* $\boldsymbol{\theta}_{W2}$.

Probability	Marginal	Mean	CI 95%
θ_{W2D}	Be(145.33, 130.67)	0.105	(0.061, 0.159)
θ_{W2H}	Be(118.33, 157.67)	0.810	(0.743, 0.870)
θ_{W2L}	Be(12.33, 263.67)	0.085	(0.045, 0.135)

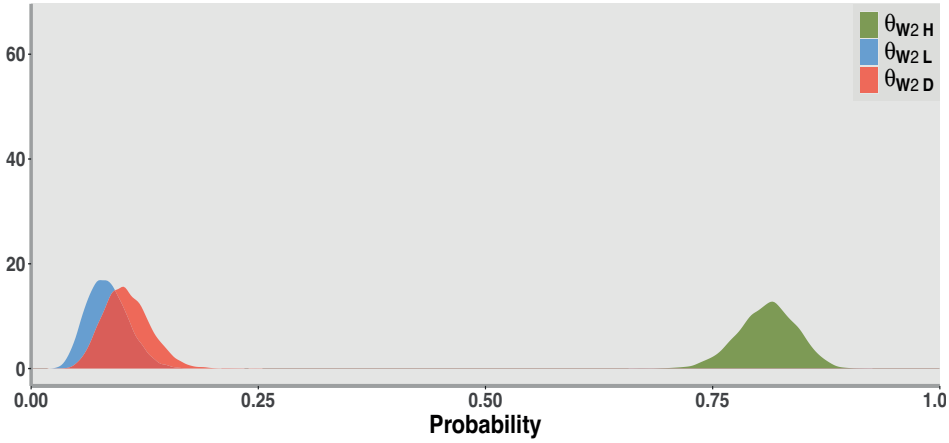

Figure 5. *Posterior distribution of the probability that a patient in ward W2 dies,* θ_{W2D}*, is sent home with a discharge,* θ_{W2H}*, or to a long-stay facility,* θ_{W2L}.

	W1	I	W2	D	H	L
Adm	0.925 [0.910; 0.938]	0.075 [0.062; 0.090]	–	–	–	–
W1	–	0.068 [0.055; 0.083]	–	0.103 [0.087; 0.121]	0.783 [0.759; 0.806]	0.046 [0.035; 0.058]
I	–	–	0.804 [0.743; 0.858]	0.196 [0.142; 0.257]	–	–
W2	–	–	–	0.105 [0.061; 0.159]	0.811 [0.743; 0.870]	0.085 [0.045; 0.135]

Figure 6. *Posterior mean and 95% credible interval associated with transition probabilities between contiguous and consecutive health states. Information relating to final states is shown in blue.*

Approximately, 81% of the patients leaving $W2$ are discharged and sent home, 10% die, and the rest, about 9%, are sent to a long-term care facility. Figure 5 shows the corresponding Beta distributions: very different in location, but with very similar variability and shape.

As a summary, Figure 6 presents the mean and a 95% credible interval of the posterior beta distributions for transition probabilities between contiguous and consecutive health states. We have chosen a matrix format because it most clearly visualises the movement of patients between different health states.

4.2. Probability of leaving hospital on discharge, dying, or being sent to a long-term care facility.

From a clinical point of view, it is important to assess the probability that a patient who enters the hospital with severe influenza will eventually be discharged home, $\theta_{A\bullet H}$, die, $\theta_{A\bullet D}$, or be sent to a long-term care facility, $\theta_{A\bullet L}$. These terminal probabilities are defined in terms of the different trajectories that connect hospital admission A to the absorbing state, D, H or L, that determine the patient's condition on discharge from hospital.

$$\begin{aligned}
\theta_{A\bullet D} &= \theta_{AIW2D} + \theta_{AID} + \theta_{AW1IW2D} + \theta_{AW1ID} + \theta_{AW1D} \\
&= \theta_{AI}\,\theta_{IW2}\,\theta_{W2D} + \theta_{AI}\,\theta_{ID} + \theta_{AW1}\,\theta_{W1I}\,\theta_{IW2}\,\theta_{W2D} \\
&\quad + \theta_{AW1}\,\theta_{W1I}\,\theta_{ID} + \theta_{AW1}\,\theta_{W1D}. \\
\theta_{A\bullet H} &= \theta_{AW1H} + \theta_{AW1IW2H} + \theta_{AIW2H} = \theta_{AW1}\,\theta_{W1H} \\
&\quad + \theta_{AW1}\,\theta_{W1I}\,\theta_{IW2}\,\theta_{W2H} + \theta_{AI}\,\theta_{IW2}\,\theta_{W2H}. \\
\theta_{A\bullet L} &= \theta_{AW1L} + \theta_{AW1IW2L} + \theta_{AIW2L} = \theta_{AW1}\,\theta_{W1L} \\
&\quad + \theta_{AW1}\,\theta_{W1I}\,\theta_{IW2}\,\theta_{W2L} + \theta_{AI}\,\theta_{IW2}\,\theta_{W2L}.
\end{aligned} \tag{5}$$

The posterior distribution for these probabilities, $\pi(\theta_{A\bullet D} \mid \mathcal{D})$, $\pi(\theta_{A\bullet H} \mid \mathcal{D})$, and $\pi(\theta_{A\bullet L} \mid \mathcal{D})$, is not analytical but, as mentioned above, it can be approximated via simulation by generating approximate samples of the posterior distribution of each direct transition probability following (3).

Figure 7 displays the posterior distribution for the probability that a patient hospitalized with severe influenza will die in the hospital, be cured, discharged and sent home, or be transferred to a long-term care unit. Posterior mean and 95% credible intervals are reported in Table 3.

Table 3. *Posterior expectation and 95% credible interval for the probability associated to the enter in the absorbing states L, H, and D.*

Probability	Mean	CI 95%
$\theta_{A\bullet L}$	0.052	(0.040, 0.065)
$\theta_{A\bullet H}$	0.814	(0.784, 0.843)
$\theta_{A\bullet D}$	0.134	(0.116, 0.155)

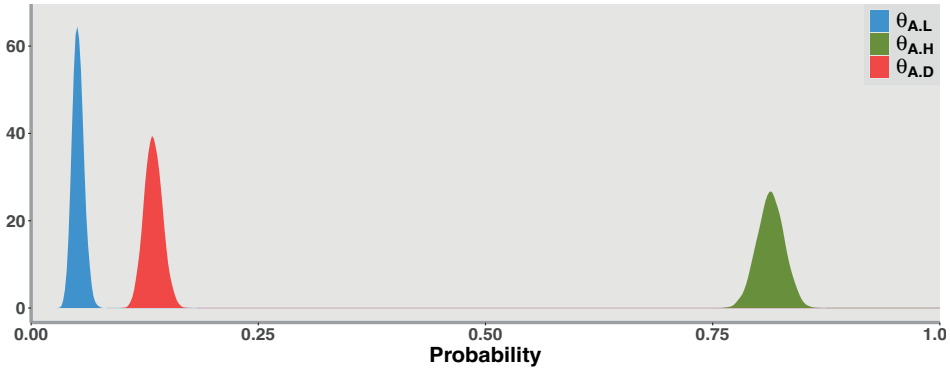

Figure 7. *Posterior distribution for the probability of dying in hospital. being sent to a long-term institution or being discharged cured and sent home.*

About 81% of patients who are hospitalised due to severe influenza will be discharged cured and sent home, about 13% will die and about 5% will be sent to a long-term care facility. As the credible intervals above and the posterior distributions for $\theta_{A\bullet L}$, $\theta_{A\bullet H}$, and $\theta_{A\bullet D}$ in Figure 7 indicate, the uncertainty associated with each of these estimates is quite small, especially that associated with dying in hospital or being sent to a long-term institution.

As a kind of general summary, Figure 8 presents, overlays on the different stages of health, the posterior mean and a 95% credibility interval for the probability that, after being admitted to hospital, a patient will visit each node of the hospital. For example, the ICU state has an approximate probability of 0.138. This probability will be that of passing through the ICU, either directly when she/he is admitted or through $W1$, that is $\theta_{A\bullet I} = \theta_{AI} + \theta_{AW1I}$.

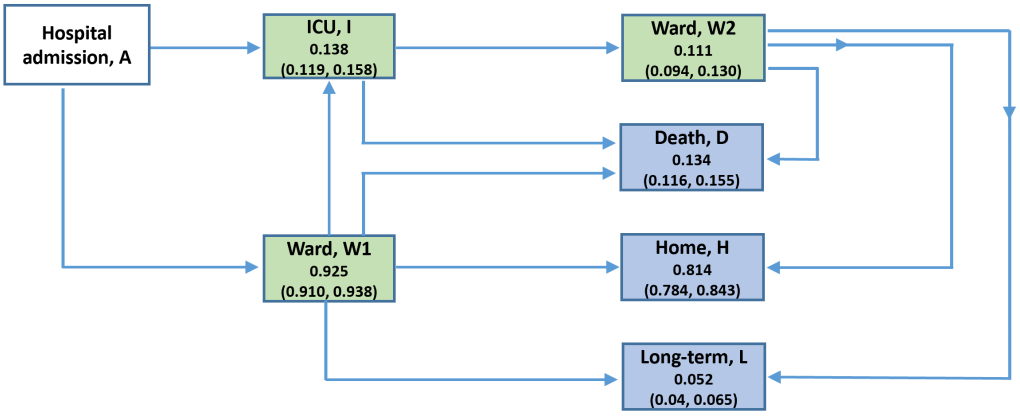

Figure 8. *Posterior mean and 95% credible interval for the probabilities associated with visiting each node from hospitalization.*

As relevant information that we have not mentioned previously, we would like to point out that visiting $W2$ means leaving the ICU alive, which occurs with an approximate probability of 0.11. We would also like to note that approximately 14% of hospitalised patients pass through the ICU.

4.3. Probability that a patient who has died, or has been discharged and sent home or has been sent to a long-term institution has spent time in the ICU

In our study, it may be interesting to simulate the posterior distribution of some inverse probabilities, such as the probability that a patient who died in the hospital, was sent to a long-term care facility or was discharged cured had previously been in the ICU, $\theta_{D\bullet I}$, $\theta_{L\bullet I}$, or $\theta_{H\bullet I}$. We start with the posterior distribution for $\theta_{D\bullet I}$. Following (4)

$$\theta_{D\bullet I} = \frac{\theta_{I\bullet D}\,\theta_I}{\theta_D},$$

where θ_D (θ_I) is the probability that a hospitalized patient dies in hospital (enters the UCI) which we have previously represented as $\theta_{A\bullet D}$ ($\theta_{A\bullet I}$), with

$$\theta_{I\bullet D} = \theta_{ID} + \theta_{IW2D} = \theta_{ID} + \theta_{IW2}\,\theta_{W2D},$$
$$\theta_{A\bullet I} = \theta_{AI} + \theta_{AW1I} = \theta_{AI} + \theta_{AW1}\,\theta_{W1I},$$

and $\theta_{A\bullet D}$ expressed in (5) in terms of transition probabilities between contiguous states. This procedure can be also followed to simulate the posterior distribution of the rest of inverse probabilities, $\theta_{L\bullet I}$ and $\theta_{H\bullet I}$.

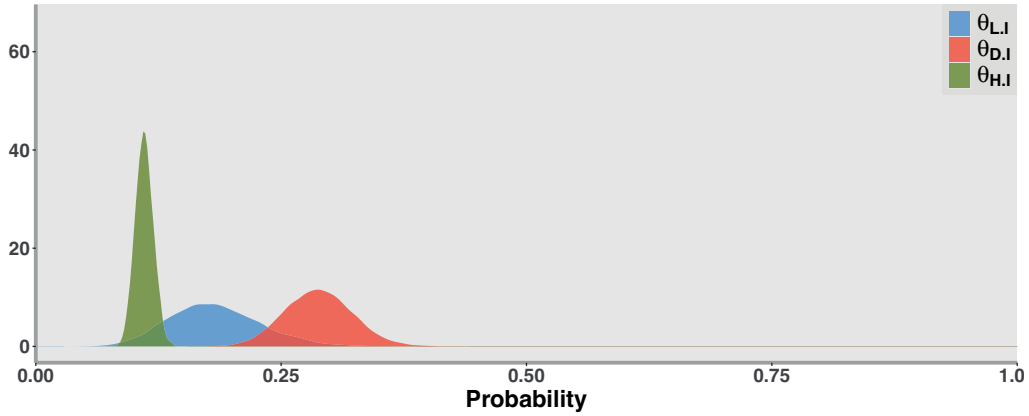

Figure 9. *Posterior distribution for the probability that a patient who has died, or has been discharged and sent home or has been sent to a long-term institution has spent time in ICU.*

Figure 9 shows the posterior distribution of the three previous probabilities. We can see that people who were discharged cured from the hospital are the least likely to have

spent time in the ICU, followed by patients who were sent to a long-stay ward, and finally people who died in hospital. We can also observe that the distribution associated with leaving the hospital cured is the one with the least uncertainty. The corresponding posterior means and 95% credible intervals are reported in Table 4.

Table 4. *Posterior mean and 95% credible interval for the probability that a patient has been in ICU given that she/he has been finally L, H, or D.*

Probability	Mean	CI 95%
$\theta_{L\bullet I}$	0.183	(0.100, 0.280)
$\theta_{H\bullet I}$	0.111	(0.093, 0.128)
$\theta_{D\bullet I}$	0.288	(0.224, 0.357)

5. Introducing covariates

The probabilities associated with transitions between states will surely depend on the different characteristics of the patients in the study population. Therefore, introducing covariates into the statistical modelling will improve the quality and accuracy of the analysis. In our case, including covariates in the model means entering the world of logistic or multinomial regression depending on the number of states, two or more than two respectively, to which a given state can visit in a single transition. In the general case where states $\{1,\ldots,J\}$ are accesible directly from state i with probabilities $\boldsymbol{\theta}_i = (\theta_{i1},\ldots,\theta_{iJ})'$ and $\boldsymbol{x}_i = (x_{i1},\ldots,x_{iQ})'$ is a vector of covariates related to state i, a multinomial regression model states that

$$\log\Big(\frac{\theta_{ij}}{\theta_{i1}}\Big) = \beta_{0ij} + \sum_{q=1}^{Q}\beta_{qij}x_{iq} = \eta_{ij},\ \ j = 2,\ldots,J, \tag{6}$$

where β_{qij} are the regression coefficients of the model and $\sum_{j=1}^{J}\theta_{ij} = 1$. This modelling implies that:

$$\begin{aligned}\theta_{i1} &= 1/(1+\textstyle\sum_{j=2}^{J}\mathrm{e}^{\eta_{ij}}),\\ \theta_{ij} &= \mathrm{e}^{\eta_{ij}}/(1+\textstyle\sum_{j=2}^{J}\mathrm{e}^{\eta_{ij}}),\ j = 2,\ldots,J.\end{aligned}$$

To complete the Bayesian model, it is required to construct a prior distribution for the model parameters, in this case the regression coefficients. We assume an inferential scenario of prior independence and very little initial information. Since we are working on the logarithmic scale, with positive and negative real values, we choose normal prior distributions centred on zero and with a wide variance. When the number of destination states from the generic i is 2, we are in the framework of logistic regression, with a logit link function for the probability associated with one of the two states directly accessible from i.

Table 5. *Number of patients who end up in the absorbing states D, H, and L from the ICU based on their age and gender.*

	Women		Men	
	Age<60	Age≥60	Age<60	Age≥60
Death, D	1	14	8	27
Home, H	25	23	28	42
Long-stay, L	1	5	2	4

To illustrate the potential of introducing covariates into the PIDIRAC study, we look at the transition that the process can make from the ICU (I) to Home (H), Death (D) or a long-stay faciliy (L) and consider two covariates relevant to the study: gender and age, the latter of which we only have in binary form, under 60 years of age or 60 years of age or older. It is worth noting that we know that ward $W2$ is an intermediate state between a person leaving the ICU without dying and the three previous states. For the sake of simplicity, we will omit this in the example, considering only moving from the ICU to the three aforementioned absorbing states. Table 5 shows the number of patients who ended up in H, L and D from the ICU in relation to their age and sex. We can see that category L has very few observations, which seems too little to feed inferences in this group.

It is well known that people who are sent to long-term care facilities often have very serious health problems with a high probability of death. That is why we have combined categories D and L into a single category, which we will represent as DL. We analyze these data by means of a logistic regression model for the probability associated to the DL category in the logit scale:

$$\text{logit}(\theta_{I\,\cdot\,DL}) = \beta_0 + \beta_1 I_{(Man)} + \beta_2 I_{(Age>=60)}. \tag{7}$$

This model implies the probability $\theta_{I\,\cdot\,DL} = \mathrm{e}^{\beta_0}$ in the group of women under the age of 60, $\theta_{I\,\cdot\,DL} = \mathrm{e}^{\beta_0+\beta_2}$ for women aged 60 or over, $\theta_{I\,\cdot\,DL} = \mathrm{e}^{\beta_0+\beta_1}$ in men less than 60 and $\theta_{I\,\cdot\,DL} = \mathrm{e}^{\beta_0+\beta_1+\beta_2}$ in men aged 60 or more.

Bayesian inference completes this model with a prior distribution on the model parameters, $\pi(\beta_0, \beta_1, \beta_2)$. We deal with an environment of prior independence and little initial information, in particular

$$\pi(\beta_0, \beta_1, \beta_2) = \pi(\beta_0)\pi(\beta_1)\pi(\beta_2),$$

with $\pi(\beta_i) = \mathcal{N}(0, 10^2)$. The posterior distribution $\pi(\beta_0, \beta_1, \beta_2 \mid \mathcal{D})$ was approximated by using Markov chain Monte Carlo (MCMC) methods through JAGS Software (Plummer, 2003). The MCMC algorithm ran for three Markov chains with 200000 iterations after a burn-in period with 20000. The effective iterations were thinned by storing every 30th iteration in order to decrease autocorrelation in the sample. Convergence of the chains to the posterior distribution is assessed through the potential scale

reduction factor, *Rhat*, and the effective number of independent simulation draws, n_{eff}. In all cases, the *Rhat* values are equal or near 1 and $n_{eff} > 100$, thus indicating that the distribution of the simulated values between and within the three chains is practically identical and also that sufficient MCMC samples have been obtained, respectively.

Figure 10 shows the approximate posterior distribution for the probability that a person in $W2$ will die, be discharged and sent home, or be referred to a long-stay service, by gender and age.

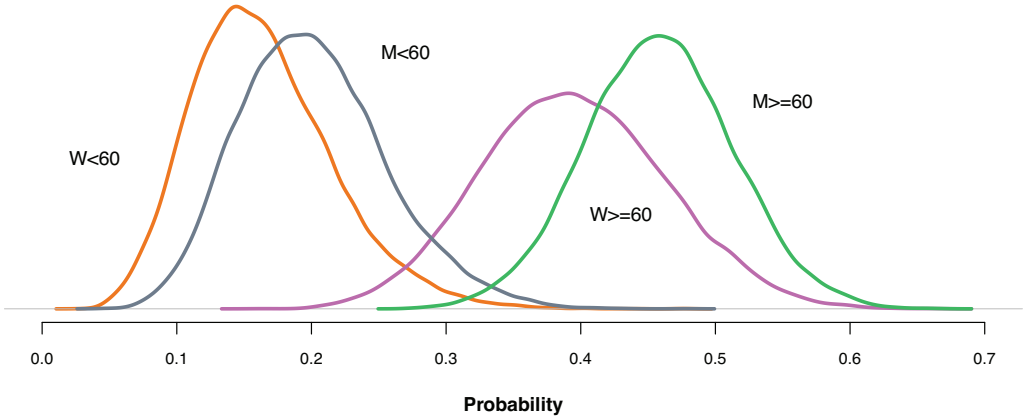

Figure 10. *Posterior distribution for the probability of dying or being sent to a long-term care facility after ICU in women under 60 years (W<60) or aged 60 or over (W>=60), and men in the same age categories (M<60 and M>=60).*

Table 6. *Number of patients who end up in the absorbing states D, H, and L from the ICU based on their age and gender.*

Group	Mean	Sd	CI 95%
$W < 60$	0.163	0.052	(0.078, 0.279)
$W >= 60$	0.395	0.068	(0.266, 0.532)
$M < 60$	0.200	0.054	(0.107, 0.320)
$M >= 60$	0.457	0.055	(0.352, 0.567)

It is interesting to note that age and gender area very relevant factors in the possible recovery of patients hospitalised with severe influenza. In the group of people under 60, the approximate probability of dying or being sent to a care home is approximately 0.163 for women and 0.20 for men. These probabilities increase dramatically in people aged 60 or over: 0.395 for women and 0.457 for men.

6. Conclusions

Probabilistic DAGs are very helpful representations of complex environments with stochastic dependencies. Bayesian inference in DAGS defined by random events is a powerful framework for understanding and assessing the prevalence and uncertainty associated with the different trajectories in the system.

The graphical representation of the evolution of patients admitted to hospital as a consequence of severe influenza through a DAG and the subsequent statistical analysis provides valuable clinical information on the severity of the disease as well as on the utilisation of healthcare resources. This information is key to hospital resource planning because it helps identify the human, material and financial resources needed to ensure quality and efficient hospital care.

Our work provides a flexible framework that allows the inclusion of potentially relevant additional information in terms of demographic (sex, age) or clinical (comorbidities) covariates, as well as other types of information such as length of stay in the different services, or even considering the potential variability between the different hospitals participating in the study. These last two proposals would correspond to topics in multistate models and hierarchical Bayesian models, respectively.

The application of Bayesian methods usually requires the use of complex computational tools because the underlying posterior distribution is not analytical. In our case, this is not so, because the inferential process associated with multinomial probabilities is completely analytical: the Dirichlet distribution is the conjugate family with respect to the multinomial probabilistic model. This scenario facilitates the simplicity and transparency of the implementation of the procedures involved. Furthermore, the use of basic procedures for generating observations from the resulting posterior distributions allows us to obtain very good approximations of non-analytical posterior distributions, as is the case with probabilities associated with non-adjacent paths or inverse transitions. However, when we introduce complexity into the modelling, such as the inclusion of covariates, we need to use intensive computational procedures, MCMC methods in our case.

Acknowledgements

The authors are very grateful to two anonymous reviewers, whose valuable comments and suggestions have contributed significantly to improving the quality of the paper. This paper is partially supported by the project PID2023-148158OB-I00, funded by Ministerio de Ciencia, Innovación y Universidades (Spain) and the project PID2022-136455NB-I00, funded by Ministerio de Ciencia, Innovación y Universidades (Spain) (MCIN/AEI/10.13039 /501100011033/FEDER, UE) and the European Regional Development Fund.

References

Acosta, L., Soldevila, N., Torner, N., Martínez, A., Ayneto, X., Rius, C., Jané, M., Domínguez, A. and the Influenza Surveillance Network of Catalonia, PIDIRAC. (2021). Influenza Vaccine Effectiveness in Preventing Severe Outcomes in Patients Hospitalized with Laboratory-Confirmed Influenza during the 2017-2018 Season. A Retrospective Cohort Study in Catalonia (Spain). *Viruses* 13, 14655.

Alvares, D., Armero, C., and Forte A. (2018). What does objective mean in a Dirichlet-Multinomial process? *International Statistical Review (ISI)* 86(1), 106-18.

Armero, C., García-Donato, G., Jiménez-Puerto, J., Pardo-Gordó, S., and Bernabeu, J. (2021). Bayesian classification for dating archaeological sites via projectile points. *SORT* 45(I), 33–46.

Barber, D. (2012). *Bayesian reasoning and machine learning.* Cambridge University Press.

Berger, J. O., Bernardo, J. M. and Sun, D. (2015). Overall objective priors. *Bayesian Analysis*, 10(1), 189-221.

Carrillo-Santisteve, P., Ciancio, B. C., Nicoll, A., & Luigi Lopalco, P. (2012). The importance of influenza prevention for public health. *Human Vaccines & Immunotherapeutics*, 8(1), 89-95.

Congdon, P. (2005). *Bayesian Models for categorical Data*. WILEY.

Cowell, R. G., Dawid, A., Lauritzen, S., Spiegelhalter, D. (1999), *Probabilistic Networks and Expert Systems*. Springer.

European Centre for Disease Prevention and Control (ECDC). Disease facts about seasonal influenza. Available online: https://www.ecdc.europa.eu/en/seasonal-influenza/facts [last access: March 2025].

Langer, J., Welch, V.L., Moran, M.M., Cane, A., Lopez, S.M., Srivastava, A., Enstone, A.L., Sears, A., Markus, K.J, Heuser, M., Kewley, R.M., and Whittle, I.J. (2023). High Clinical Burden of Influenza Disease in Adults Aged $\geq$ 65 Years: Can We Do Better? A Systematic Literature Review. *Adv Ther* 40(I), 1601–1627. https://doi.org/10.1007/s12325-023-02432-1

Macias, A. E., McElhaney, J. E., Chaves, S. S., Nealon, J., Nunes, M. C., Samson, S. I., Seet, B. T., Weinke, T., and Yu, H. (2021). The disease burden of influenza beyond respiratory illness. *Vaccine*, 39, Supplement 1, A6-A14.

Paget, J., Iuliano, A. D., Taylor, R. J., Simonsen, L., Viboud, C., Spreeuwenberg, P. (2022). Estimates of mortality associated with seasonal influenza for the European Union from the GLaMOR project. *Vaccine*, 40(9), 1361-1369,

Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *J. Inst. Actuaries*, 73(2), 285-334.

Pietrasik, T. (2023). Influenza (Seasonal). *World Health Organization (WHO)*. Available online: https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal) [last access: March 2025].

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (DSC 2003), Vienna, 20-22 March 2003, 1-10.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Snacken, R., and Brown C. (2015). New developments of influenza surveillance in Europe. *Eurosurveillance* 20(4), pii=21020. Available online: https://doi.org/10.2807/ese.20.04.21020-en.

Soldevila, N., Acosta, L., Martínez, A., Godoy, P., Torner, N., Rius, C., Jané, M., Domínguez, A. & the Surveillance of Hospitalized Cases of Severe Influenza in Catalonia Working Group. (2021). Behavior of hospitalized severe influenza cases according to the outcome variable in Catalonia, Spain, during the 2017-2018 season. *Scientific Reports* 11:13587.

Information for authors

Author Guidelines

SORT accepts for publication only original articles that have not been submitted simultaneously to any other journal in the areas of statistics, operations research, official statistics or biometrics. Furthermore, once a paper is accepted it must not be published elsewhere in the same or similar form.

SORT is an **Open Access** journal which **does not** charge publication **fees**.

Articles should be preferably of an applied nature and may include computational or educational elements. Publication will be exclusively in English. All articles will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board.

Submission of papers must be in electronic form only at our **RACO** (Revistes Catalanes en Accés Obert) submission site. Initial submission of the paper should be a single document in **PDF** format, including all **figures and tables** embedded in the main text body. **Supplementary material** may be submitted by the authors at the time of submission of a paper by uploading it with the main paper at our RACO submission site. **New authors**: please register. Upon successful registration you will be sent an e-mail with instructions to verify your registration.

The article should be prepared in **double-spaced** format, using a **12-point** typeface. SORT strongly recommends the use of its LaTeX template.

The **title page** must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (75–100 words) followed by the keywords and MSC2010 Classification of the American Mathematical Society.

Before submitting an article, the author(s) would be well advised to ensure that the text uses **correct English**. Otherwise the article may be returned for language improvement before entering the review process. The article must also use inclusive language, that is, language that avoids the use of certain expressions or words that might be considered to exclude particular groups of people, especially gender-specific words, such as “man”, “mankind”, and masculine pronouns, the use of which might be considered to exclude women. The article should also inform about whether the original data of the research takes gender into account, in order to allow the identification of possible differences.

Bibliographic references within the text must follow one of these formats, depending on the way they are cited: author surname followed by the year of publication in parentheses [e.g., Mahalanobis (1936) or Rao (1982b)])]; or author surname and year in parentheses, without comma [e.g. (Mahalanobis 1936) or (Rao 1982b) or (Mahalanobis 1936, Rao 1982b)]. The complete reference citations should be listed alphabetically at the end of the article, with multiple publications by a single author listed chronologically. Examples of reference formats are as follows:

- Article: Casella, G. and Robert, C. (1998). Post-processing accept-reject samples: recycling and rescaling. Journal of Computational and Graphical Statistics, 7, 139–157.

- Book: Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). Bayesian Data Analysis, 2nd Ed. Chapman & Hall / CRC, New York.

- Chapter in book: Engelmann, B. (2006). Measures of a rating’s discriminative power-applications and limitations. In: Engelmann, B. and Rauhmeier, R. (eds), The Basel Risk Parameters: Estimation, Validation, and Stress Testing. Springer, New York.

- Online article (put issue or page numbers and last accessed date): Marek, M. and Lesaffre, E. (2011). Hierarchical generalized linear models: The R package HGLMMM. Journal of Statistical Software, 39 (13). http://www.jstatsoft.org/v39/i13. Last accessed 28 March 2011.

Explanatory footnotes should be used only when absolutely necessary. They should be numbered sequentially and placed at the bottom of the corresponding page. **Tables and figures** should also be numbered sequentially.

Papers should not normally exceed about **25 pages** of the **PDF** format (**40 pages** of the format provided by the SORT **LaTeX** template) including all figures, tables and references. Authors should consider transferring content such as long tables and supporting methodological details to the online supplementary material on the journal's web site, particularly if the paper is long.

Authors must indicate the **source of funding** for the articles.

Authors can **preprint** their manuscripts during the submission process and showcase their work to the global research community, before it is accepted or published. This can be done in any non-commercial preprint server such as **archiv**.

Once the article has positively passed the first review round, the executive editor assigned with the evaluation of the paper will send comments and suggestions to the authors to improve the paper. At this stage, the executive editor will ask the authors to submit a revised version of the paper using the SORT **LaTeX** template.

Once the article has been accepted, the journal editorial office will **contact the authors** with further instructions about this final version, asking for the source files.

The Library of Catalonia (depending on the Government of Catalonia) periodically records SORT website and stores it indefinitely in the repository **PADICAT** (Patrimoni Digital de Catalunya: https://www.padicat.cat/)

Submission Preparation Checklist

As part of the submission process, authors are required to check off their submission's compliance with all of the following items, and submissions may be returned to authors that do not adhere to these guidelines.

1. The submitted manuscript follows the guidelines to authors published by SORT
2. Published articles are under a Creative Commons License BY-NC-ND
3. Font size is 12 point
4. Text is double-spaced
5. Title page includes title, name(s) of author(s), professional affiliation(s), complete address of corresponding author
6. Abstract is 75-100 words and contains no notation, no references and no abbreviations
7. Keywords and MSC2010 classification have been provided
8. Bibliographic references are according to SORT's prescribed format
9. English spelling and grammar have been checked
10. Manuscript is submitted in PDF format

Creative Commons License

Copyright notice and author opinions

Authors transfer the exploitation rights of their works to the journal. The Institut d'Estadística de Catalunya holds the copyright ownership of the contents published in the journal. Authors may deposit a copy of their works in repositories, as specified in the self-archiving policy. Published articles represent the author's opinions; the journal SORT does not necessarily agree with the opinions expressed in the published articles.

Self-archiving policy

The journal SORT allows the deposit and dissemination of the published version of the contributions in any institutional, subject and/or multidisciplinary repository. The repository must contain the information about the publication in the journal and the corresponding link. The journal allows the deposit and dissemination of article preprints, but recommends being linked to the published version.

Statement of ethics and good practices

As a journal co-edited by the Universitat de Barcelona, SORT declares that follows its "Declaration of Ethics and Good Practices for Scientific Journals" (https://diposit.ub.edu/dspace/handle/2445/97665)

SORT Statistics and Operations Research Transactions
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58 - 08003 Barcelona. SPAIN
Tel. +34-93.557.30.76
sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.