SORT

Statistics and Operations Research Transactions

Volume
48

Number 1, January-June 2024

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Statistics and Operations Research Transactions

Volume 48, Number 1, January-June 2024

eISSN: 2013-8830

www.idescat.cat/sort/

Aims

SORT (*Statistics and Operations Research Transactions*) —formerly Qüestiió— is an international journal launched in 2003 and distributed in printed form as well as in digital form online. From 2024 it will be published in digital form only. It is published twice-yearly by the Institut d'Estadística de Catalunya (Idescat), co-edited by the Universitat Politècnica de Catalunya, Universitat de Barcelona, Universitat Autònoma de Barcelona, Universitat de Girona, Universitat Pompeu Fabra, Universitat de Lleida i Universitat Rovira i Virgili and the cooperation of the Spanish Section of the International Biometric Society, the Catalan Statistical Society and the Departament de Recerca i Universitats, of the Generalitat de Catalunya. SORT promotes the publication of original articles of a methodological or applied nature or motivated by an applied problem in statistics, operations research, official statistics or biometrics as well as book reviews. We encourage authors to include an example of a real data set in their manuscripts. SORT is an Open Access journal which does not charge publication fees.

SORT is indexed and abstracted in the *Science Citation Index Expanded* and in the *Journal Citation Reports* (Clarivate Analytics) from January 2008. The journal is also described in the *Encyclopedia of Statistical Sciences* and indexed as well by: *Current Index to Statistics*, *Índice Español de Ciencia y Tecnología*, *MathSci*, *Current Mathematical Publications and Mathematical Reviews*, and *Scopus*.

SORT represents the third series of the *Quaderns d'Estadística i Investigació Operativa* (*Qüestiió*), published by Idescat since 1992 until 2002, which in turn followed from the first series *Quaderns d'Estadística, Sistemes, Informàtica i Investigació Operativa* (1977-1992). The three series of *Qüestiió* have their origin in the *Cuadernos de Estadística Aplicada e Investigación Operativa*, published by the UPC till 1977.

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/

FECYT-073/2023
Fecha de certificación: 20 de mayo de 2011 (2ª convocatoria)
Válido hasta: 28 de julio de 2024

eISSN: 2013-8830
SORT 48 (1) January-June (2024)

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 48
Number 1
January-June 2024
eISSN: 2013-8830

Invited article *(with discussion)*

Articles

SORT 48 (1) January-June 2024, 3-66 DOI: 10.57645/20.8080.02.13

A diffusion-based spatio-temporal extension of Gaussian Matérn fields

Finn Lindgren[1], Haakon Bakka[2], David Bolin[3], Elias Krainski[3] and Håvard Rue[3,4]

Abstract

Gaussian random fields with Matérn covariance functions are popular models in spatial statistics and machine learning. In this work, we develop a spatio-temporal extension of the Gaussian Matérn fields formulated as solutions to a stochastic partial differential equation. The spatially stationary subset of the models have marginal spatial Matérn covariances, and the model also extends to Whittle-Matérn fields on curved manifolds, and to more general non-stationary fields. In addition to the parameters of the spatial dependence (variance, smoothness, and practical correlation range) it additionally has parameters controlling the practical correlation range in time, the smoothness in time, and the type of non-separability of the spatio-temporal covariance. Through the separability parameter, the model also allows for separable covariance functions. We provide a sparse representation based on a finite element approximation, that is well suited for statistical inference and which is implemented in the R-INLA software. The flexibility of the model is illustrated in an application to spatio-temporal modeling of global temperature data.

Keywords: *Stochastic partial differential equations, Diffusion, Gaussian fields, Non-separable space-time models, INLA, finite element methods.*

[1] School of Mathematics, The University of Edinburgh, Scotland. As part of the EUSTACE project, Finn Lindgren received funding from the European Union's Horizon 2020 Programme for Research and Innovation, under Grant Agreement no. 640171.

[2] Kontali, Oslo, Norway.

[3] CEMSE Division, King Abdullah University of Science and Technology, Kingdom of Saudi Arabia.

[4] Address for correspondance: Professor Håvard Rue, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia. Email: haavard.rue@kaust.edu.sa
Received: June 2023

1. Introduction

1.1. Modelling spatio-temporal data

Statistical models for spatio-temporal data have applications in areas ranging from the analysis of environmental data (Cameletti et al., 2013) and climate data (Wood et al., 2004; Fuglstad and Castruccio, 2020), to resource and risk modeling (e.g., of wildfires, Serra et al. (2014)), disease modeling (Bhatt et al., 2015; Moraga, 2019), and ecology (Yuan et al., 2017; Zuur, Ieno and Saveliev, 2017). These models typically use spatio-temporal random effects, defined as Gaussian spatio-temporal stochastic processes and rely on a large body of theoretical and methodological literature (Stein, 2012; Gelfand et al., 2010; Cressie and Wikle, 2011, and references therein).

At best, this theory is carefully studied when the spatio-temporal model is constructed, so that the model with the most appropriate assumptions can be used. In practice, however, users of statistical software often choose a model based on convenience. If there are available code examples, the choices made in these will often be carried forward into future analyses. For example, users of R-INLA (Rue, Martino and Chopin, 2009, 2017; van Niekerk et al., 2021; van Niekerk and Rue, 2024; Gaedke-Merzhäuser et al., 2022; van Niekerk et al., 2023) construct space-time models through Kronecker products of a spatial Matérn model, and first- or second-order autoregressive models in time, following the code examples in Krainski et al. (2019). This paper is aimed at improving the general practice of space-time data analysis, by providing a new family of spatio-temporal stochastic processes for use as random effects in statistical software.

We will mainly discuss stochastic processes $u(\boldsymbol{s},t)$ that are stationary and spatially isotropic, i.e., the covariance function can be written as $\text{cov}(u(\boldsymbol{s}_1,t_1),u(\boldsymbol{s}_2,t_2)) = R(h_s,h_t)$, where $h_s = ||\boldsymbol{s}_1 - \boldsymbol{s}_2||$ and $h_t = |t_1 - t_2|$, but will also extend these process models to spatial non-stationarity and processes on general manifolds. We consider these stochastic processes in the context of hierarchical models, as a latent model component, observed through some measurement process, with no direct measurements of the stochastic process itself. Consider, for example, a model with a linear predictor

$$\eta(\boldsymbol{s},t) = \sum_{i=1}^{m} X_i(\boldsymbol{s},t)\beta_i + f_1\{z_1(\boldsymbol{s},t)\} + \ldots + f_k\{z_k(\boldsymbol{s},t)\} + u(\boldsymbol{s},t), \tag{1}$$

that is connected to the response y through some likelihood or loss function (Bissiri, Holmes and Walker, 2016) such that $\text{E}\{y(\boldsymbol{s},t)\} = g\{\eta(\boldsymbol{s},t)\}$ for some fixed and known function g. Here X_i and z_j are covariates that vary over both space and time, β_i the regression coefficient for the fixed effects, and $f_j(z_j)$ are random effects. Typical examples are splines and latent Gaussian processes used to approximate the effect of altitude or distance to coastline. This common situation with a stochastic process as a model component impacts the methodological considerations we make. The predictor is also a spatio-temporal stochastic process, with a covariance function that can be deduced from the assumptions on the model components. However, properties of the predictor that we

may discover by investigating the covariance function of the predictor may not be shared by the spatio-temporal model component u because of the other factors. Hence, we may have little prior information about the covariance structure of the spatio-temporal model component, except that it should be physically realistic, and should mimic the dependency structure in models of physical processes.

Users of software for spatio-temporal modelling most often use separable models (see, e.g., Bakka et al. (2018); Krainski et al. (2019)), i.e., models where u has a covariance function of the form $R(h_s,h_t) = R_s(h_s)R_t(h_t)$, for some spatial and temporal marginal covariance functions $R_s(\cdot)$ and $R_t(\cdot)$. This is typically not because this is a desired property, but since such models are readily available in statistical software, and there are many good arguments for why models should not be assumed separable, see Stein (2005), Cressie and Huang (1999), Fonseca and Steel (2011), Rodrigues and Diggle (2010), Gneiting (2002), Sigrist, Künsch and Stahel (2015), Wikle (2015).

1.2. The Matérn family of covariance functions

The most well known family of covariance functions for stationary random fields on $\mathbb{R}^d$ is the Matérn covariance,

$$R_M(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa h)^\nu K_\nu (\kappa h), \tag{2}$$

where $\nu, \kappa > 0$ are smoothness and scale parameters, σ^2 is the variance of the corresponding random field, K_ν is the Bessel function of the second kind of order ν, and Γ is the Gamma function. An important property of this covariance family is that it allows for explicit control of the differentiability of the corresponding stochastic process through the parameter ν. It further allows for control of the practical correlation range $r = \sqrt{8\nu}/\kappa$ (Lindgren et al., 2011). The covariance function is usually attributed to Matérn (1960), and it was advocated early by Handcock and Stein (1993) and Stein (2012). See Guttorp and Gneiting (2006) for a historical account of the covariance function and its connections to various areas in physics.

The goal of this paper is to extend the Matérn covariance function to a family of spatio-temporal covariance functions. One way of doing this would be to extend the covariance function to a spatio-temporal covariance. However, we argue that it is better to base the extension on some of the other equivalent mathematical representations, or *views*, of Gaussian Matérn fields. One such alternative representation is the stochastic partial differential equation (SPDE) representation by Whittle (1963). Specifically, a Gaussian Matérn field on $\mathbb{R}^d$ solves the SPDE

$$(\kappa^2 - \Delta)^{\alpha/2} u = \mathcal{W}, \tag{3}$$

where $\kappa > 0$, Δ is the Laplacian, $\mathcal{W}$ is Gaussian white noise, and $\alpha = \nu + d/2$. Via the SPDE representation, we note that a Gaussian Matérn field has precision operator $Q = (\kappa^2 - \Delta)^\alpha$. The precision operator (as well as the pseudo-differential operator $(\kappa^2 -$

$\Delta)^{\alpha/2}$) are defined in terms of Fourier transforms (Lindgren et al., 2011), and informally, we get the Fourier transform of the precision operator by replacing derivatives with d-dimensional wave-numbers $\boldsymbol{w}$. For any precision operator which is a polynomial in the Laplacian, $Q = \text{Poly}(-\Delta)$, such as the Matérn operator with $\alpha \in \mathbb{N}$, this results in a polynomial $\mathcal{F}(Q) = \text{Poly}(\|\boldsymbol{w}\|^2)$. This function is the reciprocal of the spectrum of the Gaussian process, illustrating why many common spectrums are the reciprocal of an even polynomial. In fact, Rozanov (1977) showed that a stationary stochastic process on $\mathbb{R}^d$ is Markov if and only if the spectral density is the reciprocal of a polynomial, and more generally, a stochastic process is Markov if the precision operator is a local operator, which is the case for integer powers of the Laplacian. For further details on the theory of the SPDE representation, see Kelbert, Leonenko and Ruiz-Medina (2005); Prévôt and Röckner (2007); Lindgren et al. (2011); Bolin and Kirchner (2020); Bolin et al. (2023).

We could also represent a Gaussian Matérn field as a stochastic integral with respect to white noise. For Gaussian Matérn fields, the kernel in the integral representation is the Green's function of the differential operator (see, e.g., Bolin, 2014). This representation can be used to define other valid covariance functions by replacing the Green's function with some other kernel (see, e.g., Fuentes, 2002; Higdon, 2002; Rodrigues and Diggle, 2010).

The modeling approaches stemming from these different views of the Gaussian Matérn fields can be thought of as implicit and explicit. In implicit approaches such as the covariance-based representation, one does not have a direct formulation of the process itself, and properties of interest need to be derived from the covariance function. In explicit, or constructive, approaches one directly defines the process through, e.g., an SPDE or a stochastic integral with the desired properties encoded. In this paper we follow the explicit approach to construct a stochastic process based on diffusion processes. Other properties, such as covariance non-separability, are then merely consequences of the explicit construction.

1.3. SPDE-based spatio-temporal generalisations of the Matérn covariance family

There is a large literature on spatio-temporal covariance models see, e.g., Porcu, Furrer and Nychka (2021) and the references within). Broadly, models for spatio-temporal Gaussian random fields can be divided into two categories; the implicit second-order covariance based models and explicit dynamical models (Cressie and Wikle, 2011; Roques, Allard and Soubeyrand, 2022). It should be noted that Porcu et al. (2021), contrary to this terminology, classifies the SPDE-based methods as implicit since they do not explicitly specify the covariance function. However, the covariance functions is merely a property of the process, and only indirectly defines the process family, whereas dynamical models directly determine the spatial and temporal evolution of the process. As shown by Lindgren et al. (2011), the covariance does not have an inherent advantage over spectral and precision operator/matrix methods, for practical applications and computations.

The second-order model specifications specify the Gaussian process properties by specifying its first two moments, and are thus based on formulating valid spatio-temporal covariance functions. In dynamical model specifications, the evolution of the Gaussian process is explicitly described either by specifying the conditional distributions of the current state of the process given its past through conditional distributions (e.g., Storvik, Frigessi and Hirst, 2002), or by specifying the process as the solution to an SPDE (Cressie and Wikle, 2011). One of the advantages with the dynamical approach is that it avoids the difficulties with formulating flexible and yet valid spatio-temporal covariance functions that can possess features such as non-separability or non-stationarity. In this work we focus on dynamical models specified through SPDEs, which makes extension to non-stationary fields and manifold models straightforward.

Several papers have been using the SPDE view to suggest models for spatio-temporal stochastic processes. A common extension of Matérn covariance fields to space-time is to use it as the spatial component in a separable model. Jones and Zhang (1997) discuss how separable covariance functions can be understood through differential operators, written as $L = L_s L_t$, where L_s is a purely spatial operator and L_t is a purely temporal operator. In agreement with Jones and Zhang (1997), we note that these operators are almost never encountered when modeling physical reality, hence, separable models are typically not physically motivated models for the spatio-temporal process.

Whittle (1963) considered a spatio-temporal stochastic process formulated as a solution to

$$\frac{\partial u}{\partial t} + (\kappa^2 - \Delta) u(\boldsymbol{s},t) = \varepsilon(\boldsymbol{s},t), \tag{4}$$

where $\varepsilon(\boldsymbol{s},t)$ is a stationary spatio-temporal noise process. Whittle (1986) denoted the model as a "diffusion-injection model" since it is a diffusion process with stochastic variability "injected" through the noise process on the right-hand side. Despite being a natural spatio-temporal extension of the Matérn model (3) with $\alpha = 2$, the model does not have any flexibility in terms of differentiability in space or time. Jones and Zhang (1997) proposed a generalization, with greater flexibility for the marginal spatial covariances, by considering the fractional SPDE

$$\left(\frac{\partial}{\partial t} + \left(\kappa^2 - \Delta\right)^{\alpha/2}\right) u(\boldsymbol{s},t) = \mathrm{d}\mathcal{E}(\boldsymbol{s},t), \tag{5}$$

where $\mathrm{d}\mathcal{E}$ is space-time Gaussian white noise. When requiring spatial operator order $\alpha > d$, this SPDE has regular continous solutions. In order to allow smaller operator orders α, such as a dampened ordinary diffusion operator with $\alpha = 2$ on $\mathbb{R}^2$, as in Whittle (1963), the driving noise process would need to have spatial dependence. We will make this precise in later sections. An advantage with (5) is that the spatial smoothness can be controlled, since the solutions on the spatial domain $\mathbb{R}^d$ have smoothness $\nu_s = \alpha - d/2$. The disadvantage is that the temporal smoothness also is determined by α. As we will see later, the marginal temporal differentiability of the solution is $\nu_t = (1 - d/\alpha)/2$.

A model with general differentiability in both space and time was formulated by Stein (2005), who considered Gaussian spatio-temporal models specified through the spectrum

$$S(\boldsymbol{w}_s, w_t) = \{c_1(a_1^2 + \|\boldsymbol{w}_s\|^2)^{\alpha_1} + c_2(a_2^2 + |w_t|^2)^{\alpha_2}\}^{-\nu}, \tag{6}$$

where $c_1 > 0, c_2 > 0, a_1, a_2$ are scale parameters, $a_1^2 + a_2^2 > 0$, α_1, α_2 and ν are smoothness parameters with further restrictions in order to obtain a model with finite variance. For example, on a two-dimensional spatial and one-dimensional temporal domain, $2/\alpha_1 + 1/\alpha_2 < 2\nu$ is required. Stein's model can also be stated as an SPDE driven by space-time white noise,

$$\left(c_1(a_1^2 - \Delta)^{\alpha_1} + c_2\left(a_2^2 - \frac{\partial^2}{\partial t^2}\right)^{\alpha_2}\right)^{\nu/2} u(\boldsymbol{s},t) = \mathrm{d}\mathcal{E}(\boldsymbol{s},t), \tag{7}$$

see Krainski (2018) and Vergara, Allard and Desassis (2022). A related model based on spectral densities, which also has separable models as a special case, was considered by Fuentes, Chen and Davis (2008).

The case $\alpha = 2$ of (5) for general dimension was considered in (Lindgren et al., 2011, Section 3.5), suggesting the generalisation

$$\left(\frac{\partial}{\partial t} + \kappa^2 + \boldsymbol{m}\cdot\nabla - \nabla\cdot\boldsymbol{H}\nabla\right) u(\boldsymbol{s},t) = \mathrm{d}\mathcal{E}_Q(\boldsymbol{s},t), \tag{8}$$

where $\boldsymbol{H}$ is a constant diffusion matrix, $\boldsymbol{m}$ is an advection (transport) vector field, and the innovation process $\mathrm{d}\mathcal{E}_Q(\boldsymbol{s},t)$ white noise in time but is sufficiently smooth in space to generate regular solutions $u(\boldsymbol{s},t)$; see Lindgren et al. (2011) and Sigrist et al. (2015, Sec 2.2). Physically, this model might be interpreted as a dampened advection-diffusion process, with the driving mechanism of the space-time field, such as introducing new mass (or, particles) into the system, having positive spatial correlation. See also Liu, Yeo and Lu (2022); Clarotto et al. (2022).

In this work, we introduce another generalisation of the models by Jones and Zhang (1997) and Lindgren et al. (2011) that intersects, but is otherwise distinct from, the Stein model family.

1.4. Outline

In Section 2 we introduce a new family of SPDE-based spatio-temporal stochastic processes. Model properties such as spatial and temporal differentiability, and parameter interpretations, are presented in Section 3.We present a sparse basis function representation in Section 4, and an implementation in R-INLA (Rue et al., 2009) in Supplementary Materials, which allows us to construct models with different likelihoods and several random effects in a generalised additive model context. In Section 5, we present a forecasting example that illustrates clearly the difference between separable models and non-separable diffusion-based models, and an application to a global temperature dataset. The article concludes with a discussion in Section 6.

2. A diffusion-based family of spatio-temporal stochastic processes

In this section we define a diffusion-based extension of the Gaussian Matérn fields to a family of spatio-temporal stochastic processes (abbreviated DEMF). The main property we aim for is that the process should be a Gaussian Matérn field when considered for a fixed time point in $\mathbb{R}^d$. That is, when the process is considered on the spatial domain $\mathcal{D} = \mathbb{R}^d$, the spatial marginalisations of the process have Matérn covariances. When the models are considered on a general (compact) manifold $\mathcal{D}$, the spatial marginalisations are solutions to a generalised spatial Whittle-Matérn model on $\mathcal{D}$ (Lindgren, Bolin and Rue, 2022).

Consider the operator $L_s = \gamma_s^2 - \Delta$, $\gamma_s > 0$, on a spatial domain $\mathcal{D}$, including any boundary conditions needed for compact domains. Let the precision operator for the generalised Whittle-Matérn covariances be $Q(\gamma_s, \gamma_e, \alpha) = \gamma_e^2 L_s^\alpha$, corresponding to solutions $v(\boldsymbol{s})$ to the spatial stochastic SPDE

$$\gamma_e L_s^{\alpha/2} v(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathcal{D} \tag{9}$$

where $\mathcal{W}$ is a spatial white noise process, as discussed by Whittle (1963) and Lindgren et al. (2011). When $\mathcal{D} = \mathbb{R}^d$, and a stationary condition is imposed, these processes are regular Matérn processes. We then define a noise process $\mathrm{d}\mathcal{E}_Q(\boldsymbol{s},t)$ as Gaussian noise that is white in time but correlated in space, with precision operator $Q = Q(\gamma_s, \gamma_e, \alpha_e)$ for some non-negative α_e. For $a > 0$, the cumulative time-integral process

$$\mathcal{E}_Q(\boldsymbol{s}, (0,a]) = \int_{t=0}^{a} \mathrm{d}\mathcal{E}_Q(\boldsymbol{s},t) \tag{10}$$

is a Q-Wiener process (Da Prato and Zabczyk, 2014), with spatial precision operator Q/a.

The case of a separable covariance model with a Matérn covariance in space and an exponential covariance in time is obtained from the stationary solutions to

$$\left(\frac{\partial}{\partial t} + \kappa\right) u(\boldsymbol{s},t) = \mathrm{d}\mathcal{E}_Q(\boldsymbol{s},t), \qquad (\boldsymbol{s},t) \in \mathcal{D} \times \mathbb{R}. \tag{11}$$

This is a spatial generalisation of the Ornstein-Uhlenbeck processes. We aim to produce a space-time model with diffusive behaviour. For this, we replace the dampening coefficient κ in (11) with a power of the dampended diffusion operator L_s, defining a model family of the time-stationary solutions to iterated diffusion-like processes

$$\left(\gamma_t \frac{d}{dt} + L_s^{\alpha_s/2}\right)^{\alpha_t} u(\boldsymbol{s},t) = \mathrm{d}\mathcal{E}_Q(\boldsymbol{s},t), \qquad (\boldsymbol{s},t) \in \mathcal{D} \times \mathbb{R}. \tag{12}$$

When $\mathcal{D} = \mathbb{R}^d$, the space-stationary solutions are used. For compact manifolds with boundary, the operators L_s and Q are equipped with suitable boundary conditions on $\partial\mathcal{D}$.

In total, the model has three non-negative smoothness parameters $(\alpha_t, \alpha_s, \alpha_e)$ and three positive scale parameters $(\gamma_t, \gamma_s, \gamma_e)$. It is not immediately obvious how the definition (12) would be interpreted for non-integer powers α_t. However, by taking advantage of the spectral properties of the operators, we define the following model, which has an operator that more clearly allows fractional powers α_t, as

$$\left(-\gamma_t^2 \frac{d^2}{dt^2} + L_s^{\alpha_s}\right)^{\alpha_t/2} u(\boldsymbol{s},t) = \mathrm{d}\mathcal{E}_Q(\boldsymbol{s},t), \qquad (\boldsymbol{s},t) \in \mathcal{D} \times \mathbb{R}. \tag{13}$$

Theorem 2.1. *For $\mathcal{D} = \mathbb{R}^d$, as well as for other domains where L_s has well defined positive powers, the definitions* (12) *and* (13) *of the Gaussian process $u(\boldsymbol{s},t)$ coincide for $\alpha_t \in \mathbb{N}$.*

Proof. This can be seen by applying the techniques developed in Vergara et al. (2022). Alternatively, the transfer function $G(\omega_t)$ (see Lindgren, 2012, Chapter 4) for the temporal linear filter defined by the operator in (12) is $G(\omega_t) = (i\gamma_t\omega_t + L_s^{\alpha_s/2})^{\alpha_t}$, well-defined for positive integers α_t, and has $|G(\omega_t)|^2 = (\gamma_t^2\omega_t^2 + L_s^{\alpha_s})^{\alpha_t}$. The transfer function $H(\omega_t)$ for the temporal linear filter defined by the operator in (13) is $H(\omega_t) = (\gamma_t^2\omega_t^2 + L_s^{\alpha_s})^{\alpha_t/2}$, well-defined for positive α_t. We see that $|G(\omega_t)|^2 = |H(\omega_t)|^2$, so the spectral properties of the two process definitions coincide for positive integer α_t values. ■

It should be noted that it would be possible to give a more direct definition of the model (12) with fractional α_t, but this would require more sophisticated mathematical tools, which is outside the scope of this work.

The two representations make it clear that the model with $\alpha_e = 0$ is a special case of the Stein (2005) model family, with $a_1 = 0$ and $\alpha_1 = 1$ in (7), and that the model of Jones and Zhang (1997) is obtained by setting $\alpha_e = 0$ and $\alpha_t = 1$ in (12).

The use of the same spatial operator L_s in the left hand side of (13) as in the precision operator on the right hand side is what causes the spatial marginalisation of the process to be Matérn fields in the simplest case, as will be shown in Section 3. The parameters α_t, α_s, and α_e determine the differential operator orders involved in the SPDE operator and therefore also the smoothness properties of the process, as shown in Section 3.

The model can be further generalised by allowing the γ parameters to vary across space. This is most straightforward for γ_s, since that only alters the L_s operator. For complex domains, as well as when L_s is generalised to vary across space, the resulting solutions are not space-stationary, but still have marginal spatial properties defined by powers of L_s. The practical precision construction in Section 4 can be generalised to separable non-stationarity, where γ_t is allowed to depend on time and γ_s and γ_e depends on space, since that retains commutativity between the temporal and spatial operators.

2.1. Compact domains and manifolds

For compact domains, the model definitions include some form of boundary conditions. These boundary conditions induce boundary effects near the domain boundary, and as shown in Lindgren et al. (2011), if such effects are undesirable, one can extend the domain by at least the spatial range. By taking advantage of a non-stationary spatial operator, the barrier method introduced by Bakka et al. (2019) can also be used to nearly eliminate boundary effects, as well as to obtain models that appropriately take complex geography into account. Similarly, all the common extensions onto curved manifolds, such as the globe, can be implemented using the same approaches as for $\mathbb{R}^d$. This includes the finite element methods used in Section 4, but also Fourier-like spectral basis function expansions given by the eigenfunctions of the Laplacian, either given in closed form, e.g., spherical harmonics on the globe, or obtained numerically from finite element eigenfunction computations. See Lindgren et al. (2022) for an overview of the literature on these alternative methods.

3. Parameter interpretations and model properties

In this section we discuss marginal spatial and temporal properties of the diffusion-based model (12). In order to simplify the exposition, we focus on the ordinary Matérn covariance case when the spatial domain is $\mathcal{D} = \mathbb{R}^d$. In this case, the space-time spectral density of the stationary solutions $u(\boldsymbol{s},t)$ to (13) is

$$S_u(\boldsymbol{\omega}_s,\omega_t) = \frac{1}{(2\pi)^{d+1}\gamma_e^2[\gamma_t^2\omega_t^2 + (\gamma_s^2 + \|\boldsymbol{\omega}_s\|^2)^{\alpha_s}]^{\alpha_t}(\gamma_s^2 + \|\boldsymbol{\omega}_s\|^2)^{\alpha_e}}, \tag{14}$$

for $(\boldsymbol{\omega}_s,\omega_t) \in \mathbb{R}^d \times \mathbb{R}$. The space-time covariance function is given by the Fourier integral

$$R_u(\boldsymbol{s},t) = \int_{\mathbb{R}}\int_{\mathbb{R}^d} \exp[i(\boldsymbol{\omega}_s\cdot\boldsymbol{s} + \omega_t t)] S_u(\boldsymbol{\omega}_s,\omega_t)\,\mathrm{d}\boldsymbol{s}\,\mathrm{d}t \tag{15}$$

for spatial lags $\boldsymbol{s}$ and temporal lags t.

3.1. Sample path continuity and differentiability theory

For fields with Matérn covariance functions, the degree of differentiability is encoded in the smoothness index ν. For models with space-time spectral density given by (14), the marginal covariance in time is not generally of the Matérn class, so we need to use more general conditions for determining the smoothness.

The differentiability of a stationary process $x(t)$, $t \in \mathbb{R}$, is determined by the decay rate of its spectral density. If $S(\omega) \sim \omega^{-\gamma}$ for some $\gamma > 0$ for large ω, then the process is a times mean square differentiable for all $a < \frac{\gamma-1}{2}$ (Stein, 2005).

For stationary Gaussian processes, stronger statements of almost sure sample path continuity of derivatives and Hölder continuity can be made. The technical details can be found in Section 9.3 of Cramér and Leadbetter (1967) and Scheuerer (2010), and are

summarised in Appendix A, including a more formal characterisation of the smoothness index. The results show that Gaussian processes with spectral densities satisfying $S(\boldsymbol{\omega}) \sim \|\boldsymbol{\omega}\|^{-2\nu-d}$ for some $\nu > 0$ and large $\|\boldsymbol{\omega}\|$ have smoothness index ν. This means that the sample paths have almost surely continuous derivatives of order up to and including $k = \lceil \nu \rceil - 1$, and that the derivatives of order k are Hölder of index a for any $0 < a < \nu - k$. Further, the sample paths are almost surely in the Sobolev spaces $W^{b,2}$ for any $b < \nu$, on finite subsets of $\mathbb{R}^d$. Although these results are derived specifically for $\mathbb{R}^d$, it is clear that sample path properties of Whittle-Matérn fields on more general but smooth domains will have similar, and usually identical, local differentiability properties, based on the decay rate of the eigenspectrum of the Laplacian. In particular, the spectral Fourier representations on the 2D sphere $\mathbb{S}^2$ lead to series that converge under the same conditions as the continuous spectra on $\mathbb{R}^2$.

The smoothness index ν can be interpreted as the smallest value for which some form of weak continuity does *not* hold. For a process on a multidimensional domain with potentially different smoothness in different directions, Theorem A.1 and smoothness definition in Appendix A will be applied to the one-dimensional marginals of the process.

3.2. Properties of the spatio-temporal model

We can now show that the spatial marginals of $u(\boldsymbol{s},t)$, i.e. for fixed t, are Matérn covariance fields, given that the smoothness parameters are chosen appropriately. To keep some notational brevity, we first define the unit variance and range Matérn covariance function $R_\nu^M(t)$,

$$R_\nu^M(t) = \frac{1}{\Gamma(\nu)2^{\nu-1}} t^\nu K_\nu(t), \quad t \geq 0, \tag{16}$$

and the scaling constants

$$C_{\mathbb{R}^d,\alpha} = \frac{\Gamma(\alpha - d/2)}{\Gamma(\alpha)(4\pi)^{d/2}},$$

for $d = 1,2,3,\ldots$ and $\alpha > d/2$. These appear as variance scaling constants for the regular Whittle-Matérn SPDE models.

Proposition 3.1. *Define the effective spatial marginal operator order* $\alpha = \alpha_e + \alpha_s(\alpha_t - 1/2)$ *and assume that* $\alpha > d/2$. *Then the solution* $u(\boldsymbol{s},t)$ *to* (13) *has marginal spatial covariance function*

$$\operatorname{cov}(u(\boldsymbol{s}_1,t), u(\boldsymbol{s}_2,t)) = \sigma^2 R_{\nu_s}^M(\gamma_s \|\boldsymbol{s}_2 - \boldsymbol{s}_1\|)$$

where $\nu_s = \alpha - d/2$ *is the spatial smoothness index and*

$$\sigma^2 = \frac{C_{\mathbb{R},\alpha_t} C_{\mathbb{R}^d,\alpha}}{\gamma_e^2 \gamma_t \gamma_s^{2\alpha-d}}. \tag{17}$$

Proof. See Appendix D.1, that also includes a derivation of the marginal spatial cross-spectra for different time lags. ■

Proposition 3.2. *Assume $\alpha_t, \alpha_s, \alpha_e$ satisfy $\alpha > d/2$. Then the temporal smoothness index of the solutions $u(\boldsymbol{s},t)$ to* (13) *is $\nu_t = \min\left[\alpha_t - \frac{1}{2}, \frac{\nu_s}{\alpha_s}\right]$, and for $d = 2$, the marginal temporal spectrum is*

$$S_t(\omega_t) \propto {}_2F_1\left(\alpha_t, \frac{\alpha_e - 1}{\alpha_s} + \alpha_t, \frac{\alpha_e - 1}{\alpha_s} + \alpha_t + 1; -\omega_t^2 \gamma_t^2 / \gamma_s^{2\alpha_s}\right),$$

where ${}_2F_1$ denotes the hypergeometric function.

Proof. See Appendix D.2. ■

For integer values of the operator orders, the hypergeometric function can be expressed using elementary functions. When $\alpha_t = \alpha_s = 2$ and $\alpha_e = 0$ for $d = 2$, we obtain

$$S_t(\omega_t) \propto \int_0^\infty \frac{1}{(\widetilde{\omega}_t^2 + (1+\nu)^2)^2}\,\mathrm{d}\nu = \frac{\arctan(\widetilde{\omega}_t)}{2\widetilde{\omega}_t^3} - \frac{1}{2\widetilde{\omega}_t^2(\widetilde{\omega}_t^2 + 1)}, \tag{18}$$

where $\widetilde{\omega}_t = \omega_t \gamma_t / \gamma_s^{\alpha_s}$, showing that the marginal temporal covariance is not a Matérn covariance. The exception is the separable case, where the temporal covariance function is a Matérn covariance function with smoothness index $\alpha_t - 1/2$.

Corollary 3.2.1. *Assume that $\alpha_s = 0$, $\alpha_t > 1/2$, and $\alpha_e > d/2$. Then the stationary solutions $u(\boldsymbol{s},t)$ to* (13) *have a separable space-time covariance function where the spatial covariance is given by Proposition 3.1 and the marginal temporal covariance function is*

$$C(u(\boldsymbol{s},t_1), u(\boldsymbol{s},t_2)) = \sigma^2 R_{\nu_t}^M(\gamma_t^{-1}|t_2 - t_1|),$$

where $\nu_t = \alpha_t - 1/2$ and σ^2 is given by (17) *with $\alpha = \alpha_e$.*

Proof. Follows directly from the product form of the space-time spectrum (14). ■

In Table 1, we summarise the general smoothness results, as well as some important special cases. The special cases denoted *diffusion* are generalised analogues of the diffusion-injection model (4), and the special *critical diffusion* model is later used in Sections 4 and 5. The general conditions on the α parameters that give well defined solutions are encoded in the spatial and temporal smoothness conditions $\nu_s > 0$ and $\nu_t > 0$, and can also be written as the conditions $\alpha = \alpha_e + \alpha_s(\alpha_t - 1/2) > d/2$ and $\alpha_t > 1/2$.

Table 1. *Summary of the smoothness properties of the solutions $u(\boldsymbol{s},t)$ for different values of the parameters $\alpha_t, \alpha_s, \alpha_e$, together with some examples. Here ν_t and ν_s respectively denote the temporal and spatial smoothnesses of the process.*

α_t	α_s	α_e	Type	ν_t	ν_s
α_t	α_s	α_e	General	$\min\left[\alpha_t - \frac{1}{2}, \frac{\nu_s}{\alpha_s}\right]$	$\alpha_e + \alpha_s(\alpha_t - \frac{1}{2}) - \frac{d}{2}$
α_t	0	α_e	Separable	$\alpha_t - \frac{1}{2}$	$\alpha_e - \frac{d}{2}$
α_t	α_s	$\frac{d}{2}$	Critical	$\alpha_t - \frac{1}{2}$	$\alpha_s(\alpha_t - \frac{1}{2})$
α_t	α_s	0	Fully non-separable	$\alpha_t - \frac{1}{2} - \frac{d}{2\alpha_s}$	$\alpha_s(\alpha_t - \frac{1}{2}) - \frac{d}{2}$
1	2	$\alpha_e > \frac{d}{2}$	Sub-critical diffusion	$1/2$	$\alpha_e + 1 - \frac{d}{2}$
1	2	$\frac{d}{2}$	Critical diffusion	$1/2$	1
1	2	$\frac{d}{2} - 1 < \alpha_e < \frac{d}{2}$	Super-critical diffusion	$\nu_s/2$	$\alpha_e + 1 - \frac{d}{2}$
1	0	2	Separable	$1/2$	$2 - \frac{d}{2}$
3/2	2	0	Fractional diffusion	$1 - \frac{d}{4}$	$2 - \frac{d}{2}$
2	2	0	Iterated diffusion	$\frac{3}{2} - \frac{d}{4}$	$3 - \frac{d}{2}$

3.2.1. Quantifying non-separability

From Table 1 we can see that the α_e parameter controls the type of non-separability. An important case is $\alpha_e = 0$, which we refer to as fully non-separable models. The spectral density for such models is a subfamily of the Stein (2005) spectral model family. The degree of non-separability can be quantified by the relation between α_e and the effective marginal spatial operator order α. We introduce the non-separability parameter $\beta_s = 1 - \alpha_e/\alpha = 1 - \alpha_e/(\nu_s + d/2) \in [0,1]$, where $\beta_s = 0$ gives a separable model, and $\beta_s = 1$ gives a "maximally non-separable" model. Assuming given values for the temporal smoothness $\nu_t > 0$, spatial smoothness $\nu_s > 0$, and non-separability $\beta_s \in [0,1]$, we can find the corresponding values of $(\alpha_t, \alpha_s, \alpha_e)$. Let $\beta_*(\nu_s, d) = \frac{\nu_s}{\nu_s + d/2}$. Then

$$
\begin{aligned}
\alpha_t &= \nu_t \max\left(1, \frac{\beta_s}{\beta_*(\nu_s,d)}\right) + \frac{1}{2}, \\
\alpha_s &= \frac{\nu_s}{\nu_t} \min\left(\frac{\beta_s}{\beta_*(\nu_s,d)}, 1\right) = \frac{1}{\nu_t}\min\left[(\nu_s + d/2)\beta_s, \nu_s\right], \\
\alpha_e &= \frac{1-\beta_s}{\beta_*(\nu_s,d)}\nu_s = (\nu_s + d/2)(1-\beta_s).
\end{aligned}
$$

The critical branching point $\beta_s = \beta_*(\nu_s, d)$ motivates the term *critical* for such models. Models with $\beta_s < \beta_*(\nu_s, d)$ are *sub-critical* and models with $\beta_s > \beta_*(\nu_s, d)$ are *super-critical*. The critical models have $\alpha_t = \nu_t + 1/2$, $\alpha_s = \nu_s/\nu_t$, and $\alpha_e = d/2$. The diffusion models in Table 1 with $\alpha_t = 1$ and $\alpha_s = 2$ are of particular interest, as they arise from a basic heat equation.

Notably, the fully non-separable diffusion model ($\alpha_t = 1, \alpha_s = 2, \alpha_e = 0$) requires $d = 1$ to ensure $\nu_s > 0$, whereas the fully non-separable twice iterated diffusion model ($\alpha_t = 2, \alpha_s = 2, \alpha_e = 0$) is valid for $d \in \{1,2,3,4,5\}$.

3.2.2. Scale parameter interpretation

To improve the interpretability of the scale parameters, we define σ, r_s, and r_t via

$$\sigma^2 = \frac{C_{\mathbb{R},\alpha_t} C_{\mathbb{R}^d,\alpha}}{\gamma_t \gamma_e^2 \gamma_s^{2\alpha-d}} \tag{19}$$

$$r_s = \gamma_s^{-1} \sqrt{8\nu_s} \tag{20}$$

$$r_t = \gamma_t \gamma_s^{-\alpha_s} \sqrt{8(\alpha_t - 1/2)}, \tag{21}$$

where r_s is the correlation range as in Lindgren et al. (2011), giving approximately correlation of 0.13 at r_s distance in space (keeping time fixed). Similarly, r_t controls the temporal correlation range for the separable model. In the non-separable cases, it is the temporal correlation range for the evolution of the spatial eigenfunction corresponding to the smallest eigenvalue of the Laplacian, i.e. a constant function over space, evolving in time. Eigenfunctions for larger spatial eigenvalues have shorter temporal correlation range, so the combined effective range will typically be smaller than the nominal r_t value would indicate.

3.3. Examples

Table 2. *Four specific DEMF models on* $\mathbb{R}^2$ *or* $\mathbb{S}^2$.

Model	α_t	α_s	α_e	Type	ν_t	ν_s
A: DEMF(1,0,2)	1	0	2	Separable order 1	$1/2$	1
B: DEMF(1,2,1)	1	2	1	Critical diffusion	$1/2$	1
C: DEMF(2,0,2)	2	0	2	Separable order 2	$3/2$	1
D: DEMF(2,2,0)	2	2	0	Iterated diffusion	1	2

To simplify notation, we denote by DEMF$(\alpha_t, \alpha_s, \alpha_e)$ the model with given values for $(\alpha_t, \alpha_s, \alpha_e)$. The four models we will consider on $\mathbb{R}^2$ are defined in Table 2. For these, we choose γ_s, γ_e and γ_t, so that $\sigma = 1$, $r_s = 1$ and $r_t = 1$. This enables us to compare the non-Matérn behaviour of the temporal correlation to the spatial Matérn behaviour.

In general, the covariances are not available in closed form, but since the temporal covariance for each spatial frequency is of Matérn type, the spatial cross-spectra (derived in Appendix D.1) can be inverted numerically to obtain the cross-covariance. Specifically, the cross-covariance can be computed numerically with a 2D fast Fourier transform (FFT) computation for each fixed temporal lag (see Appendix B). This technique is related to the half-spectral space-time covariance models from Horrell and Stein (2017).

There, they focus on models where the temporal spectrum is known for each spatial location, $\mathcal{F}_t R(\boldsymbol{s},t) = f(\omega_t)g(\boldsymbol{s},\omega_t)$, but the theory also covers the case of known spatial spectrum for each time point, $\mathcal{F}_s R(\boldsymbol{s},t) = f(\boldsymbol{\omega}_s)g(\boldsymbol{\omega}_s,t)$, that we use here.

In Figure 1 we show the spatio-temporal covariance function for these four models, and the marginal spatial covariances are shown in Figure 2. There is a clear difference between the spatio-temporal covariances, even though the marginal spatial covariances are identical for the first three models.

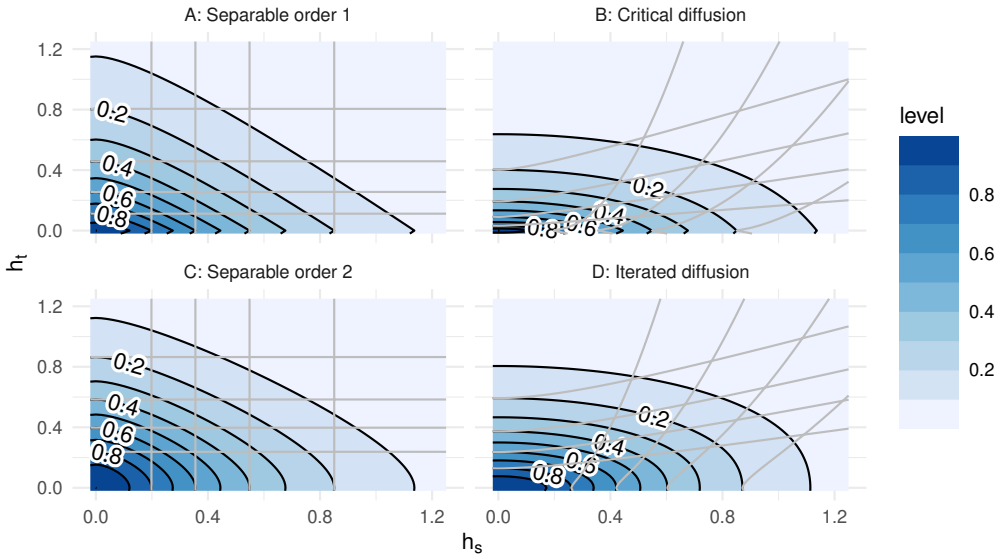


Figure 1. *The space-time covariance functions for spatial dimension $d = 2$, for the four models in Table 2, Section 3.3. The grey overlayed curves are level curves of the relative decay of the spatial and temporal covariances in relation to the marginal covariances. The non-separable models have non-orthogonal decay.*

3.4. Spheres and other manifolds

As noted earlier, the marginal spatial covariance properties of the DEMF models on general manifolds are rooted in the properties of the Whittle-Matérn operator, and depend on the specific geometry. However, the temporal structure is linked to each spatial frequency in the same way for every manifold, so we can focus on the effects on the spatial properties. Smoothness properties intuitively follow from the local properties of the differential operator on smooth manifolds, which locally behave like $\mathbb{R}^d$, so that is not the main obstacle to determining the process properties. Instead, it is the effect of the manifold's intrinsic curvature that prevents general closed form expressions for the covariance functions to be derived. On a compact manifold $\mathcal{D}$, the covariance function for models based on $L_s^{\alpha/2} = (\gamma_s^2 - \Delta)^{\alpha/2}$ (where $\alpha = \alpha_e + \alpha_s(\alpha_t - 1/2)$ in the DEMF

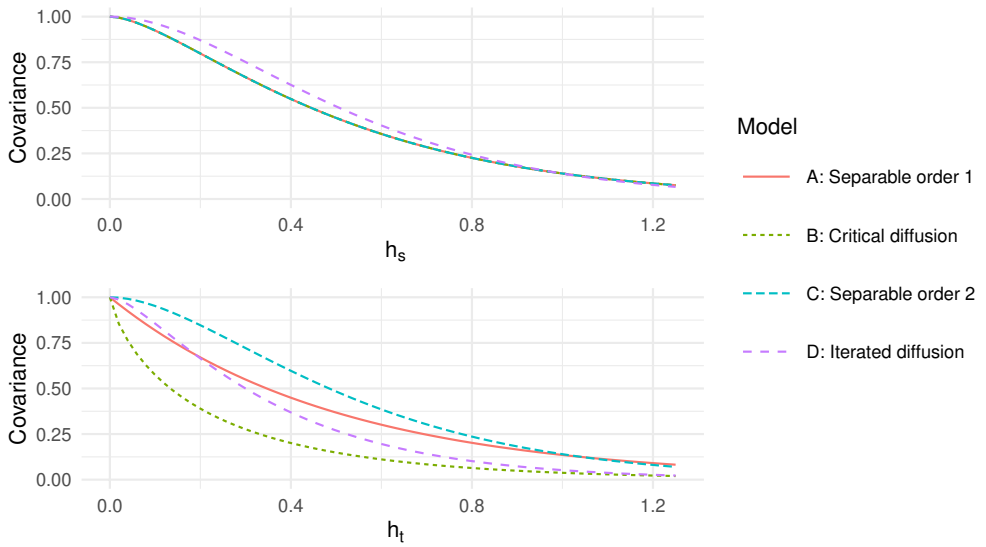


Figure 2. *The marginal spatial and temporal covariances for spatial dimension $d = 2$, for the four models in Table 2, Section 3.3. The spatial correlation is approximately 0.13 when the distance equals the range r_s. For the temporal correlations, that relationship to r_t only holds for the contribution from the evolution of a spatial constant, and the effective range has a more complex structure, depending on the combined model parameter.*

models) takes the form

$$R(\boldsymbol{s},\boldsymbol{s}') = \sum_{k=0}^{\infty} C_k \frac{1}{(\gamma_s^2+\lambda_k^2)^{\alpha}} E_k(\boldsymbol{s}) E_k(\boldsymbol{s}'), \quad \boldsymbol{s} \in \mathcal{D},$$

where (λ_k, E_k) are the eigenvalue/function pairs of the $-\nabla\cdot\nabla$ (negated Laplace-Beltrami) operator on $\mathcal{D}$, and C_k are scaling constants that depend on potential scaling of the eigenfunctions and multiplicity of eigenvalues. This was used in Lindgren et al. (2011) to show that the finite element constructions for Whittle-Matérn fields work on general manifolds. On the sphere, the eigenfunctions are the spherical harmonics, with eigenvalues $\lambda_k = k(k+1)$ with multiplicity $2k+1$. With the spherical harmonic definitions in Appendix C, the resulting covariance can be simplified to

$$R_{\mathbb{S}^2,\alpha}(\boldsymbol{s},\boldsymbol{s}';\gamma_s) = \sum_{k=0}^{\infty} \frac{2k+1}{4\pi[\gamma_s^2+k(k+1)]^{\alpha}} P_{k,0}(\boldsymbol{s}\cdot\boldsymbol{s}'), \tag{22}$$

where $P_{k,0}(\cdot)$ are Legendre polynomials of order k, and the factor $\frac{2k+1}{4\pi}$ comes from the eigenvalue multiplicity and Fourier-Bessel transform theory on the sphere (see Appendix C). It follows from the construction that the infinite series for the covariances of the process derivatives that the differentiability properties on the sphere are the same as on $\mathbb{R}^2$, as the terms $\lambda_k^a \frac{2k+1}{[\gamma_s^2+\lambda_k]^{\alpha}}$ decay at the same rate as required for the smoothness criteria on $\mathbb{R}^2$ from Appendix A.

Due to the wraparound effects on the sphere, the spatial variance contribution to the overall field variance is not the same as on $\mathbb{R}^2$, and the factor $C_{\mathbb{R}^d,\alpha}/\gamma_s^{2\alpha-d}$ in (17) needs to be replaced by a function of γ_s defined by

$$C_{\mathbb{S}^2,\alpha}(\gamma_s) = \sum_{k=0}^{\infty} \frac{2k+1}{4\pi[\gamma_s^2 + k(k+1)]^\alpha}, \tag{23}$$

obtained from the spectral representation of a spherical Whittle-Matérn field. The overall variance can then be written as $\mathrm{var}[u(\boldsymbol{s},t)] = \frac{C_{\mathbb{R},\alpha_t}}{\gamma_e^2\gamma_t} C_{\mathbb{S}^2,\alpha}(\gamma_s)$, and the asymptotic behaviour of $C_{\mathbb{S}^2,\alpha}(\gamma_s)$ as γ_s approaches 0 or ∞ is given by

$$C_{\mathbb{S}^2,\alpha}(\gamma_s) = \sum_{k=0}^{\infty} \frac{2k+1}{4\pi[\gamma_s^2 + k(k+1)]^\alpha} \sim \begin{cases} \frac{1}{4\pi\gamma_s^{2\alpha}}, & \gamma_s \to 0, \\ \frac{1}{4\pi(\alpha-1)\gamma_s^{2\alpha-2}}, & \gamma_s \to \infty. \end{cases}$$

This shows that for large γ_s, i.e. short spatial ranges, the variance of the field $u(\boldsymbol{s},t)$ on the sphere is the same as on $\mathbb{R}^2$, but for small γ_s, i.e. long spatial ranges, the spherical geometry leads to larger variance than on $\mathbb{R}^2$. For intermediate γ_s values, the upper tail of the infinite series can be bounded by tractable integrals, which also allows bounding the relative error in numerical covariance and variance evaluation, by replacing the upper series tail from $k = K$ by the integral $\int_{K+1/2}^{\infty} \frac{2k+1}{4\pi[\gamma_s^2+k(k+1)]^\alpha}\,\mathrm{d}k$. More details are given in Appendix C.2.

4. Hilbert space representation

The discussion up to this point has focused on the general continuous domain properties of the proposed model class. We will now discuss aspects of numerical implementations, suitable for inclusion in generalised additive latent Gaussian models, as available in the `INLA` and `inlabru` packages for `R`. The general construction is applicable to a wide range of basis function representations. In practice, we will use the finite element approach from Lindgren et al. (2011) due to its computational convenience, in particular in the unstructured spatial observation location and manifold domain contexts.

4.1. Hilbert space approximation

We consider general Kronecker product basis expansions

$$u(\boldsymbol{s},t) = \sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \psi_i(\boldsymbol{s})\phi_j(t)u_{ij}, \tag{24}$$

where $\{\psi_i(\boldsymbol{s}); i = 1,\dots,n_s\}$ and $\{\phi_j(t); j = 1,\dots,n_t\}$ are finite basis sets for Hilbert spaces on a spatial domain $\mathcal{D}$ and a time interval $[T_0,T_1] \subset \mathbb{R}$, respectively. We will show that projection onto the resulting Kronecker function space only involve integrals of the form $\langle\phi_j,\phi_{j'}\rangle$, $\langle(-\Delta)^{k/4}\phi_j,(-\Delta)^{k/4}\phi_{j'}\rangle$, and $\langle L_s^{k/2}\psi_i, L_s^{k/2}\psi_{i'}\rangle$. This is possible due

to the lack of interaction in the individual model operators; the operator as a whole is non-separable, but each operator term is space-time separable. This also extends to the case of a non-stationary L_s operator, as mentioned in Section 2.

Different choices of spatial and temporal basis functions have benefits and drawbacks depending on the specific modelling and data context. A natural choice for the spatial domain is local piecewise linear basis functions. Such functions were used in Lindgren et al. (2011) to construct model representations with sparse precision matrix structure for the basis expansion coefficients, via Gaussian Markov random fields (GMRF). This allows a large number of basis functions to be used, and pointwise georeferenced observations will not alter the sparseness of the posterior precision matrix, making this a versatile approach, that can also be used in combination with sparse matrix solvers developed for ordinary deterministic PDE computations. For very smooth processes, the piecewise linear basis functions can in principle be replaced by higher order local polynomials (Liu, Guillas and Lai, 2016), but this can be difficult to implement. For non-stationary $L_s = \gamma_s(\boldsymbol{s}) - \Delta$, the spatially varying $\gamma_s(\boldsymbol{s})$ values only have a local influence on the finite element construction, so the additional computational complexity lies mainly on the increased number of parameters needed to represent the spatial variation of $\gamma_s(\cdot)$.

An alternative to piecewise linear basis functions are harmonic basis functions based on the eigenfunctions of the Laplacian. These can be very efficient on domains that admit fast Fourier inversion algorithms, such as $\mathbb{R}^d$ and partially on $\mathbb{S}^2$. However, the diagonal precision matrix structure implied by the basic models is broken by scattered georeferenced observations, as the resulting posterior precision matrix becomes dense, so the utility is greatest for very smooth processes that can cut off the harmonics at a long spatial range. So-called *conditioning by kriging* can also be applied in such cases, but this is computationally expensive for large numbers of observations unless the number of basis functions is kept small. A further complication on general domains and manifolds is the lack of closed form expressions for the harmonics. Computing them with finite element methods, for example, is as expensive as applying the piecewise linear basis GMRF representations directly. They are also impractical for non-stationary operators, since the precision matrices will typically become dense instead of diagonal.

A third alternative is Karhunen-Loève expansions, which yield better approximations for fewer basis functions than harmonic basis. They can handle non-stationary operators, but need recomputing the basis for each set of parameter values, making inference expensive. For irregular data, the same problem exists of turning a sparse prior precision matrix into a dense posterior precision matrix. However, for given parameters, it can in principle be applied to the posterior distribution instead. Unfortunately, the numerical computations for each eigenfunction is at least as expensive as computing the posterior expectation using the same numerical method (e.g., finite elements) as in the GMRF computations, making the full computation much more expensive, and best suited to special cases such as computing a compact representation of a given, fixed, distribution.

Despite their practical numerical cost and other related problems, the harmonic basis and Karhunen-Loève e expansions are excellent tools for theoretical analysis, and their discrete domain formulations are essential in the theoretical proofs of the general discretisation construction below. See Lindgren et al. (2022) for further discussion on the relative merits of different basis choices.

The above considerations largely apply to the temporal basis function choice as well, with a few useful differences. First, in addition to piecewise linear basis functions, B-spline basis functions of higher order can readily be applied, and in particular second order B-splines (piecewise quadratic basis functions) provide immediate benefits with only minimal extra effort. Where piecewise linear basis functions require some form of mass lumping for operator order 2, second order B-splines can be applied with least squares finite element projection, and the resulting discretised Laplacian operator matrix has the same non-sparsity as for piecewise linear basis functions. In addition, when applied to order 1 operators, temporal interpolation in the finite-dimensional representation exhibits less quasi-deterministic fluctuations than for piecewise linear basis functions. Second, harmonic basis functions are useful for smooth cyclic processes, e.g. seasonal effects, but otherwise suffer from the same issues as in space.

4.2. Precision matrix construction

In this section we represent the stochastic processes DEMF(α_t,α_s,α_e) using general Kronecker basis Hilbert space representations. Define $u(s,t)$ on $\mathcal{D} \times \mathbb{R}$, for some polygonal domain $\Omega \subset \mathbb{R}^d$, as the solution to (13) with some boundary conditions on $\partial\mathcal{D}$. The particular choice of boundary conditions does not matter much in what follows as long as they lead to a well defined precision operator for the solutions of the equation posed on the bounded domain. However, in most practical situations one would use homogeneous Neumann boundary conditions on the spatial domain.

For implementations, we restrict the temporal domain to an interval, and we then also need to impose temporal boundary conditions. However, temporal boundary effects can be handled by direct calculations for the resulting AR(2) dependence structure for the temporal coefficients in the approximation; see Appendix E.

The projection of the solutions onto the finite Hilbert space result in a discretised model where the coefficients u_{ij} in (24) have a precision matrix that is expressed as a sum of Kronecker products. As in Lindgren et al. (2011), the approximation properties of the discretisation is directly linked to the expressiveness of the finite-dimensional Hilbert space spanned by the Kronecker basis $\{\psi_i(\boldsymbol{s})\phi_j(t), i = 1,\dots,n_s,\ j = 1,\dots,n_t\}$.

We provide the following theorem that links the continuous domain DEMF models to finite-dimensional Hilbert space representations. The theorem focuses on the link between the continuous domain precision operator and the precision matrix, necessarily assuming unique solutions with a unique covariance function. This makes it applicable, in principle, to more esoteric models involving various forms of intrinsic stationarity, i.e. non-stationary models with stationary properties with respect to some contrast filters. However, the details of such models are beyond the scope of the presentation.

Theorem 4.1. *Let $\alpha_t \in \mathbb{N}$ and consider the equation*

$$\left(-\gamma_t^2 \frac{\partial^2}{\partial t^2} + L_s^{\alpha_s}\right)^{\alpha_t/2} u(\boldsymbol{s},t) = \mathrm{d}\mathcal{E}_{\gamma_e^2 L_s^{\alpha_e}}(\boldsymbol{s},t) \quad \text{on } \mathcal{D}\times[T_0,T_1], \tag{25}$$

where $[T_0,T_1] \subset \mathbb{R}$ is a bounded interval, L_s is some spatial differential operator, and some boundary conditions on $\partial\mathcal{D}$ and at T_0 and T_1 are assumed such that the precision operator for the solutions of (25) *is well defined. Let $\{\psi_i(\boldsymbol{s}), i=1,\dots,n_s\}$ and $\{\phi_j(t), j=1,\dots,n_t\}$ be bases for finite-dimensional Hilbert spaces on $\mathcal{D}$ and $[T_0,T_1]$, respectively, chosen such that the product basis set $\{\psi_i(\boldsymbol{s})\phi_j(t),\, i=1,\dots,n_s,\, j=1,\dots,n_t\}$ form a basis for a finite-dimensional Hilbert space $V_h \subset V$, and let $u(\boldsymbol{s},t) = \sum_{i,j}\psi_i(\boldsymbol{s})\phi_j(t)u_{i,j} \in V_h$ be a finite-dimensional representation of a solution to* (25)*. Assume the following two conditions:*

(i) Let $v(t) = \sum_{j=1}^{n_t}\phi_j(t)v_j$ be a finite-dimensional approximation of a solution to

$$b^{1/2}\left(-\frac{\partial^2}{\partial t^2} + \kappa^2\right)^{\alpha_t/2} v(t) = \mathcal{W}(t), \quad \text{on } [T_0,T_1],$$

for some $b>0$, $\kappa>0$, and $\alpha_t = 1,2,\dots$, and the boundary conditions at T_0 and T_1. Assume that the precision matrix for the weights vector $\boldsymbol{v} = (v_1,\dots,v_{n_t})$ takes the form

$$b\sum_{k=0}^{2\alpha_t}\kappa^{2\alpha_t-k}\boldsymbol{J}_{\alpha_t,k/2}$$

for some symmetric matrices $\boldsymbol{J}_{\alpha_t,0}$, $\boldsymbol{J}_{\alpha_t,1/2}$, to $\boldsymbol{J}_{\alpha_t,\alpha_t}$.

(ii) Let $w(\boldsymbol{s}) = \sum_{i=1}^{n_s}\psi_i(\boldsymbol{s})w_i$ be a finite-dimensional approximation of a solution to

$$L_s^{a/2} w(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}) \quad \text{on } \mathcal{D},$$

where L_s is equipped with the boundary conditions on $\partial\mathcal{D}$, for some $a \geq 0$. Assume that the precision matrix for $\boldsymbol{w} = (w_1,\dots,w_{n_s})$ is $\boldsymbol{K}_a = \boldsymbol{C}^{1/2}\left(\boldsymbol{C}^{-1/2}\boldsymbol{K}_1\boldsymbol{C}^{-1/2}\right)^a\boldsymbol{C}^{1/2}$ for some symmetric positive definite matrix $\boldsymbol{K}_1$.

Assume additionally that the temporal precision construction in condition (i) is valid for all $\kappa \geq \lambda_0^{\alpha_s/2}/\gamma_t$, where λ_0 is the smallest eigenvalue in the generalised eigenvalue problem $\boldsymbol{K}_1\boldsymbol{e} = \boldsymbol{C}\boldsymbol{e}\lambda$. Then, the precision matrix for the collected coefficient vector $\boldsymbol{u} = (u_{1,1},u_{2,1},\dots)$ is given by

$$\boldsymbol{Q}_{\boldsymbol{u}} = \gamma_e^2\sum_{k=0}^{2\alpha_t}\gamma_t^k\boldsymbol{J}_{\alpha_t,k/2}\otimes\boldsymbol{K}_{\alpha_s(\alpha_t-k/2)+\alpha_e}.$$

Proof. The result follows from discretising the spatial dimension, diagonalising the resulting operator matrices, and applying the temporal precision structure condition to the resulting independent temporal equations. A detailed proof is given in Appendix D.3. ■

The existence of finite-dimensional representations fulfilling conditions (i) and (ii) for certain choices of basis functions follows directly from the general constructions in Lindgren et al. (2011).

For the regular Whittle-Matérn operator $L_s = \gamma_s^2 - \Delta$ on $\mathcal{D}$ we have $\boldsymbol{K}_1 = \gamma_s^2 \boldsymbol{C} + \boldsymbol{G}$. For triangulated domains with local piecewise linear basis functions with $\sum_{i=1}^{n_s} \psi_i(\boldsymbol{s}) \equiv 1$ on $\mathcal{D}$, we can take $\boldsymbol{C}$ to be the diagonal mass lumped mass matrix with $C_{i,i} = \langle \psi_i, 1 \rangle$ and symmetric sparse structure matrix $\boldsymbol{G}$ with $G_{i,j} = \langle \nabla\psi_i, \nabla\psi_j \rangle$. For domains where the orthogonal harmonic eigenfunctions of Δ are available, such as rectangular subdomains of $\mathbb{R}^d$ and spherical harmonics on $\mathbb{S}^2$, the full mass and structure matrices $\boldsymbol{C}$ and $\boldsymbol{G}$ are both diagonal, with $C_{i,i} = \langle \psi_i, \psi_i \rangle$.

In the temporal case, the same technique applies, but higher order B-spline basis functions are more easily applied, allowing, e.g., second order B-splines to be used without mass lumping. For temporal Neumann boundary conditions, $\boldsymbol{J}_{\alpha_t,k/2} = \boldsymbol{0}$ for odd $k = 1, 3, \ldots, 2\alpha_t - 1$ and $[\boldsymbol{J}_{\alpha_t,k/2}]_{i,j} = \langle (-\Delta)^{k/4}\phi_i, (-\Delta)^{k/4}\phi_j \rangle$ (or non-conformal approximations for non-smooth basis functions) for even $k = 0, 2, \ldots, 2\alpha_t$. Lemma E.1 in Appendix E can be used for first and second order B-spline basis functions for $\alpha_t = 1$ and 2 to provide approximate stationary boundary conditions by modifying the $\boldsymbol{J}_{\alpha_t,k/2}$ matrices for $k = 0, 1, \ldots, 2\alpha_t$. When such temporal boundary corrections are used, fractional orders appear in $\boldsymbol{K}_{\alpha_s(\alpha_t - k/2) + \alpha_e}$ for odd k unless α_s is an even integer. For the spatial piecewise linear finite element constructions, this would break sparsity, but for orthogonal harmonic function representations, $\boldsymbol{K}_a$ is diagonal for all $a \geq 0$, allowing the fractional powers to be used without loss of the diagonal property.

In the proof of Theorem 4.1, we see that it is sufficient that the initial temporal precision structure is valid for $\kappa \geq \kappa_0 = \gamma_s^{\alpha_s/2}/\gamma_t$. By taking a Taylor expansion for the boundary precision elements with respect to κ and κ_0, the approximation would be improved, compared with taking the Taylor expansion at $\kappa = 0$, as the expansion would be closer to the exact expression for a wider range of relevant temporal frequencies. This improvement would however come at the expense of making the matrix constructions dependent on the γ_s and γ_t parameters directly.

5. Applications

5.1. Separable vs non-separable forecasting

The difference between using separable and the non-separable models is most clearly seen when doing forecasting. To illustrate this, we simulated spatial data for time $t = 0$, and compute the posterior conditional expectation for $t = 0$, 1, and 2. For the simulation,

we used the four Matérn models defined in Table 2 with one percent (standard deviation) nugget effect added. The parameters were set to $r_s = 4.0$, $r_t = 2.5$ for the separable models and $r_t = 4.5$ for the non-separable models, and $\sigma = 1$. The scaling difference for r_t in the non-separable models compensates for the difference in parameter interpretation illustrated in Section 3.3. In the estimation, the nugget precision and the temporal range parameters r_t were kept fixed, so that only the marginal standard deviation σ and the spatial range parameter r_s were estimated for each model.

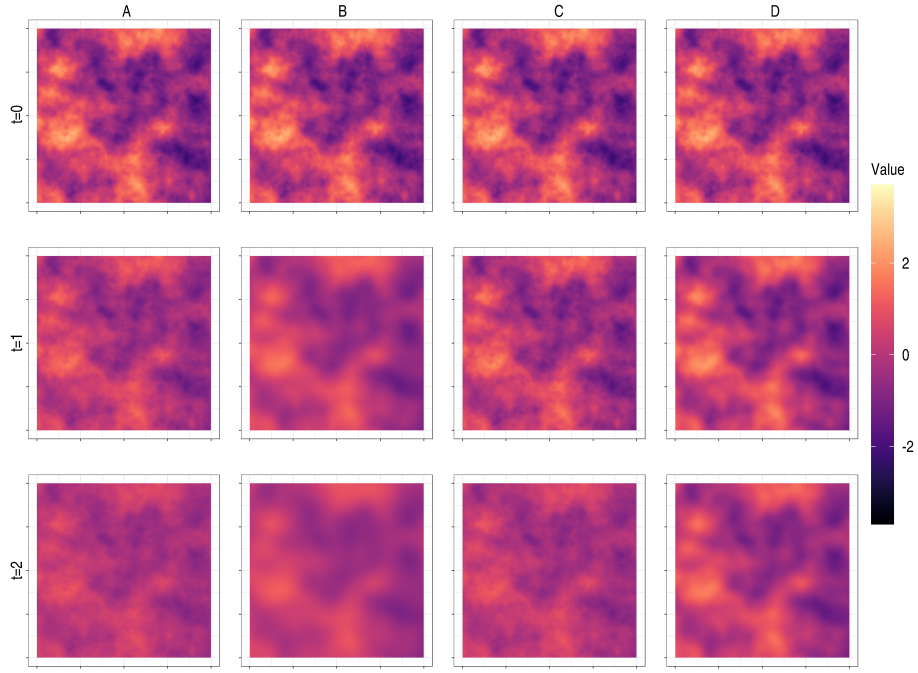


Figure 3. *Predictions from each model (A, B, C and D) when conditioned on a spatially dense dataset at $t = 0$, and no observations for $t = 1$ and $t = 2$.*

Figure 3 displays the predictions from the four models in Table 2. For $t = 0$, the results are similar for the four models, due to the highly informative data. For the predictions for $t = 1$ and $t = 2$, we see how the separable model A and C only reduce the fields point-wise towards zero, and that non-separable models B and D, exhibit spatial diffusion, as expected. This behaviour was part of the theoretical motivation of Whittle (1954, 1963), and also a major motivation for developing the DEMF family. It is also noteworthy that since the forecasts are conditional expectations based on a finite set of observations, they are smoother than the process realisations. For the separable models, this effect is not visible, since there this effect only appears on smaller spatial scales than shown, but it is clearly visible for the non-separable model. In all four cases, the posterior process realisations however have their ordinary, lower, smoothness. This is important to take into account when considering probabilistic forecasts, in particular for prediction of non-linear functionals of the process.

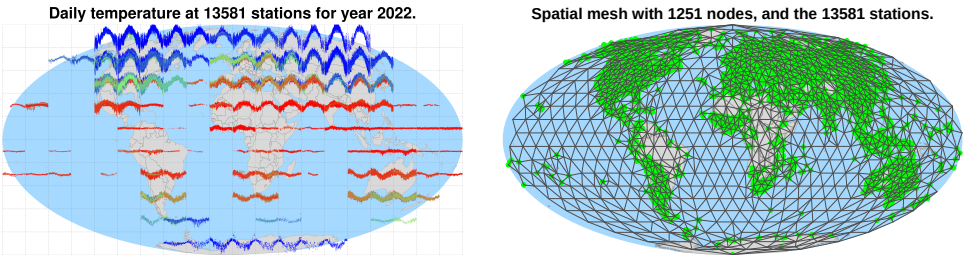


Figure 4. *Daily average temperature time series shown grouped near the corresponding locations (left), with colours based on the year average, for each station, from blue (cold), to red (warm). The locations (green) and the mesh used for the spatio-temporal model components v and u (right). The computational mesh is defined by spherical triangles directly on the globe surface, and triangles crossing the* $\pm 180^\circ$ *longitude curve are not shown in this Mollweide projection.*

5.2. Global temperature dataset

This section presents some results analysing daily temperature data, where all the code for the data cleaning, model fitting and plots are included in the supplementary material.

5.2.1. Data and model structure

We used daily data for year 2022, using minimum (TMIN) and maximum (TMAX) daily temperatures, as described in Menne et al. (2012). We cleaned the data for inconsistencies before the analysis. In particular, values beyond 7 standard deviations from the mean were treated as missing. We computed the mean of these two variables for each day at each one of 13 567 stations world-wide, a total of 4 951 955 data entries. Figure 4 (top) shows this data as time series grouped by location.

The model includes an overall level μ, the elevation in kilometres $E(\boldsymbol{s})$, a smoothed deviation from the overall mean jointly over latitude and time $b(\boldsymbol{s},t)$, a spatio-temporal random field $v(\boldsymbol{s},t)$ varying slowly in time, and a spatio-temporal random field, $u(\boldsymbol{s},t)$, capturing the daily variability. The $b(\boldsymbol{s},t)$ function is allowed to vary by latitude and time, but is fixed to zero at the equator. The linear predictor expression is

$$\eta(\boldsymbol{s},t) = \mu + \alpha E(\boldsymbol{s}) + b(\boldsymbol{s},t) + v(\boldsymbol{s},t) + u(\boldsymbol{s},t). \tag{26}$$

Each observation y_i is modelled with additive Gaussian noise with a common variance parameter, σ_e^2, so that $y_i = \eta(\boldsymbol{s}_i,t_i) + e_i$, where $(\boldsymbol{s}_i,t_i)$ is observation i, $i = 1,\ldots n$, and $e_i \sim \mathrm{N}(0,\sigma_e^2)$.

5.2.2. Model discretisation and estimation

For the $b(\boldsymbol{s},t)$ and $v(\boldsymbol{s},t)$ functions in the predictor expression (26), we defined temporal basis functions 1, $\cos[(t-1)\cdot 2\pi/365]$, and $\sin[(t-1)\cdot 2\pi/365]$. For $b(\boldsymbol{s},t)$, these were

multiplied with two quadratic basis function in $\sin(\text{latitude} \cdot \pi/180)$, which guarantees smooth behaviour with respect to the location, $\boldsymbol{s}$, at the two poles, giving a total of six basis functions. For $v(\boldsymbol{s},t)$, each of the three temporal basis functions were instead multiplied by stationary spatial Whittle-Matérn fields over the sphere forming a model term that captures the seasonal local deviation from the basic seasonal pattern described by $b(\boldsymbol{s},t)$.

The reported results were estimated using a spatial mesh with 1 251 nodes (median node distance $\sim$ 587km), shown in Figure 4, both for the spatial coefficients in v and for u. For u, we discretised the time domain with first order basis functions with one knot per day. This setting gives a spatio-temporal model for u with size 456 615. In vector form we have

$$\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{E}\alpha + \boldsymbol{B}\boldsymbol{b} + \boldsymbol{A}_v\boldsymbol{v} + \boldsymbol{A}_u\boldsymbol{u} + \boldsymbol{e},$$

where $\boldsymbol{B}$ is a six-column matrix of the evaluated basis functions for $b(\boldsymbol{s},t)$ at the observation locations and times, and $\boldsymbol{A}_v$ and $\boldsymbol{A}_u$ contains the evaluated basis functions, respectively, for $v(\boldsymbol{s},t)$ and $u(\boldsymbol{s},t)$. The vectors $\boldsymbol{b}$, $\boldsymbol{v}$, and $\boldsymbol{u}$ contain the corresponding basis weights.

We used independent priors for all the model parameters. We used a flat prior for μ and a Gaussian with mean zero and variance 100 for α and each element in $\boldsymbol{b}$. The three spatial fields in $\boldsymbol{v}$ are assumed as independent realizations each one modelled using Eq. 3 with a common spatial range r_v, and common marginal variance σ_v^2. The $u(\boldsymbol{s},t)$ term is a spatio-temporal field using one of the four models in Table 2. In total, we have six variance/range parameters to estimate. We used penalized complexity priors for all these parameters (Simpson et al., 2017; Fuglstad et al., 2018), applied to the marginal properties of the models.

To define the PC-prior for σ_e we used $\Pr(\sigma_e \geq 5) = 0.01$ and the same for σ_v and σ. We used $\Pr(r_v \leq 1000\text{km}) = 0.01$ for r_v, $\Pr(r_s \leq 600\text{km}) = 0.01$ for r_s. For r_t we used $\Pr(r_t \leq 1 \text{ days}) = 0.01$ in models A and C and $\Pr(r_t \leq 2 \text{ days}) = 0.01$ in models B and D.

5.2.3. Model fitting results

Attributing the relative contributions to each model component is non-trivial due to the posterior correlation between the components. However, a basic linear model variance decomposition, SQT = $\sum_i (y_i - \bar{y})^2$ and SQR = $\sum_i (y_i - \mathrm{E}(\eta_i \mid \boldsymbol{y}))^2$, can be obtained to define $R^2 = 1 - \text{SQR/SQT}$. We have that the predictor model $\eta(\boldsymbol{s},t)$ captures 97.18% of the variability with model B. Table 3 reports DIC, WAIC, and goodness-of-fit statistics for within-sample and leave-one-out assessment (leave-one-out log predictive density score, LCPO, see Held, Schrodle and Rue (2010), for each of the five fitted models. For within-sample assessment, R^2, mean squared error (MSE), and mean absolute error (MAE) assess the posterior mean and median only, whereas the log predictive density score (LPO), CRPS, and SCRPS assess the full predictive distribution (Gneiting et al., 2005; Bolin and Wallin, 2023). The model M_0 includes the fixed effects and $v(\boldsymbol{s},t)$,

whereas models A, B, C and D all include $\boldsymbol{u}$, using the four models in Table 2. When considering R^2, LPO, MSE, MAE, CRPS and SCRPS model B performed a slightly better. When considering DIC, WAIC and LCPO model C was slightly better.

Table 3. *Summary statistics for each estimated model. The LCPO score is the average negated log-predictive density for leave-one-out predictions, and LOP is its within-sample version. The MAE, MSE, CRSP, SCRPS were all computed as within-sample scores based on a Gaussian approximation of the posterior predictive distribution for each data point.*

Model	M_0	A	B	C	D
R^2	0.8649	0.9718	**0.9718**	0.9718	0.9718
DIC	5.8103	4.3217	4.3222	**4.3212**	4.3222
WAIC	5.8101	4.3123	4.3128	**4.3118**	4.3128
LPO	2.9046	2.1332	**2.1330**	2.1335	2.1331
LCPO	2.9051	2.1572	2.1574	**2.1569**	2.1574
MSE	19.5151	4.1163	**4.1136**	4.1193	4.1155
MAE	3.3236	1.4667	**1.4657**	1.4678	1.4661
CRPS	2.4288	1.0947	**1.0943**	1.0952	1.0945
SCRPS	1.7905	1.3947	**1.3945**	1.3949	1.3946

Table 4. *The posterior mean and standard deviation (in brackets) for each of the model parameters.*

	A	B	C	D
σ_e	2.06 (0.001)	2.06 (0.001)	2.06 (0.001)	2.06 (0.001)
r_v	5402 (217)	15304 (10683)	5242 (266)	5346 (486)
σ_v	8.5 (0.4)	20.9 (14.0)	7.9 (0.4)	8.3 (0.8)
r_s	1322 (6)	2079 (19)	1329 (7)	1365 (6)
r_t	5.60 (0.04)	42.07 (0.91)	3.92 (0.02)	6.94 (0.05)
σ	2.73 (0.01)	3.67 (0.03)	2.63 (0.01)	2.80 (0.01)

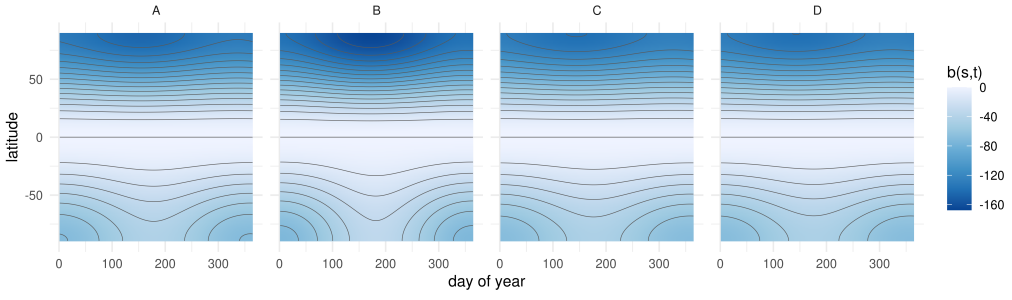


Figure 5. *The posterior mean of the smoothed seasonal latitude effect $b(\boldsymbol{s},t)$ for each model.*

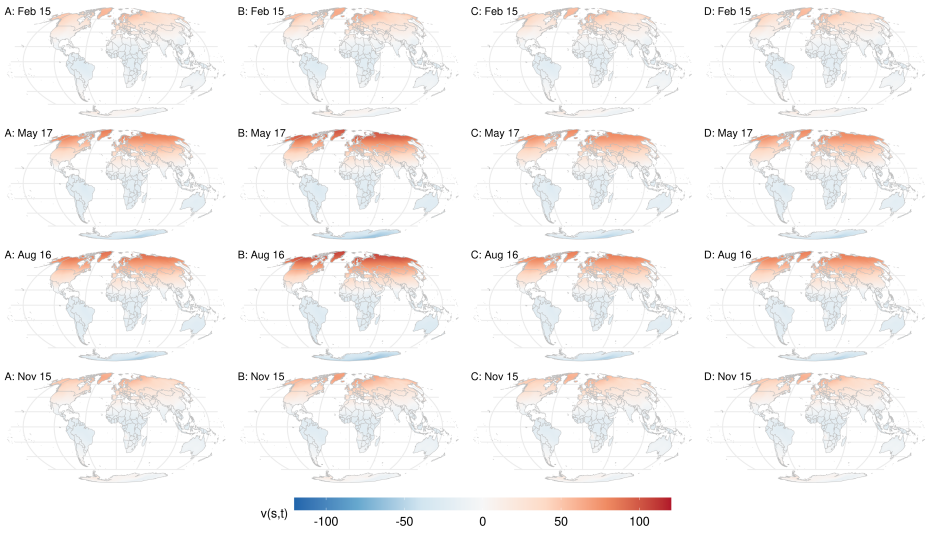


Figure 6. *The posterior mean for $v(\boldsymbol{s},t)$, at some time points, for each model. From left to right, models A, B, C, and D. From top to bottom, the time points are 46, 137, 228, and 319, corresponding to the day of year in 2022 as labelled at the top left of each plot.*

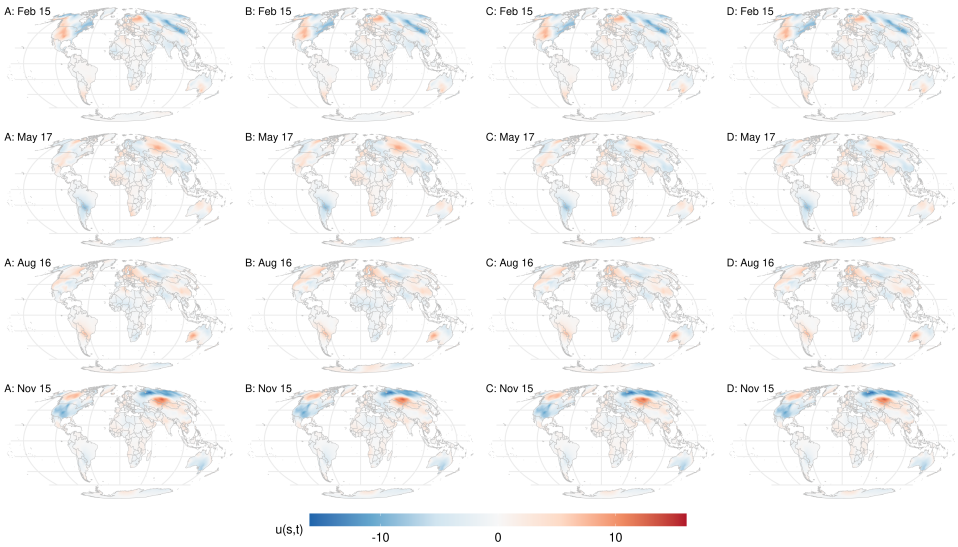


Figure 7. *The posterior mean for $u(\boldsymbol{s},t)$, at some time points, for each model. From left to right, models A, B, C, and D. From top to bottom, the time points are 46, 137, 228, and 319, corresponding to the day of year in 2022 as labelled at the top left of each plot.*

For B (using model B for $u(\boldsymbol{s},t)$), the posterior mean for μ is 34.74 and for α is -4.70. The posterior mean of $b(\boldsymbol{s},t)$ for all four models are shown in Figure 5, which displays the temperature over time and latitude. The seasonal pattern is clear, with summer and winter temperatures in the two hemispheres standing out, with in particular lower temperatures (blue) in each hemisphere's respective winter. The model range/variance parameter estimates are summarised in Table 4. For the posterior mean, we have $\mathrm{E}(r_s \mid \boldsymbol{y}) = 2078.66$ km and $\mathrm{E}(r_t \mid \boldsymbol{y}) = 42.07$ days for model B, and smaller for the other models. In particular the temporal range estimates are much smaller for the other models, ranging from 4 to 7 days, which is more realistic. These values can be interpreted through Figure 2. The posterior mean for the spatio-temporal field $v(\boldsymbol{s},t)$ for some days in 2022 is shown in Figure 6. This term is intended to capture slowly temporally varying spatial variation from the overall mean, elevation effect and the basic seasonal latitude parts of the model. It is clear that the desired interpretation is confounded with the interpretation of $b(\boldsymbol{s},t)$, shown in Figure 5. The posterior mean for the spatio-temporal field $u(\boldsymbol{s},t)$ for some days in 2022 is shown in Figure 7. This term captures the remaining spatio-temporal variation of the temperature field around the other parts of the model.

5.2.4. Forecast evaluation

As was already apparent from the diagnostic scores in Table 3, despite the temporal range parameters being different for the four models, particularly for model B, they are nearly indistinguishable with respect to direct and leave-one-out prediction distributions. Since the space-time non-separability effect is unclear in the leave-one-out setting, we extend the assessment by computing multi-horizon temporal predictions. We used the first 14 days of the data from each month to predict the following 7 days. These forecasts were done while keeping the covariance parameters and the long term spatio-temporal components $b(\boldsymbol{s},t)$ and $v(\boldsymbol{s},t)$ fixed to their posterior modes from the full joint model estimates, so that only the short-term spatio-temporal field $u(\boldsymbol{s},t)$ was reestimated for each scenario. This generated forecasts for each model for 12 different weather and seasonal conditions over the year.

Figure 8(top) shows the mean absolute error (MAE), mean squared error (MSE), mean Dawid-Sebastiani (DS, equivalent to log-score for Gaussian predictions, see Gneiting et al., 2005), and mean SCRPS summarized for each prediction horizon (1–7 days) for each of the 12 scenarios. Figure 8(bottom) shows the difference between the scores for each model to those of model D, to more clearly highlight the differences between the models. The prediction errors all exhibit increasing variability for longer forecast horizons, as well as a generally increasing trend, that mostly levels off around 6 days, which is compatible with the estimated temporal correlation length parameter r_t for models A, C, and D. For 1-day ahead forecasts, model D achieved the lowest scores, and it appears more stable than the other models for longer forecast horizons. Model B has large score variability, and is doing worse than the other three models for long forecast horizons, in particular for the scores that take forecast uncertainty into account. For more details see

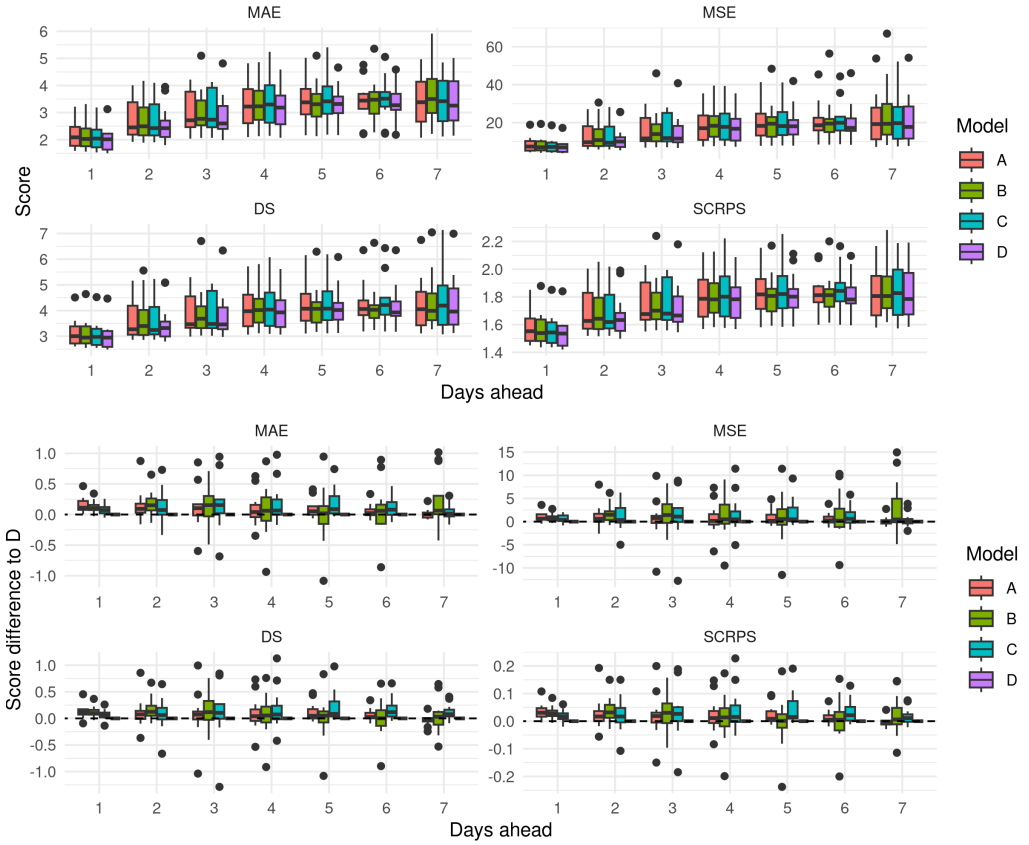


Figure 8. *Top: Multi-horizon (1–7 days) and multi scenario (one for each month of year 2022) forecast scores for predicting one week ahead. Lower scores indicate a better forecast. Bottom: The differences in scores compared with model D.*

Appendix F, where one can see that the scores are generally worse in the start and end of the year, indicating an unmodelled aspect of seasonality, e.g. in weather variability.

6. Discussion

We have developed a spatio-temporal extension of the Gaussian Matérn fields based on a fractional and stochastic version of the physical diffusion equation considered by Whittle (1954, 1963). We named the new family the diffusion-based extension of the Matérn field (DEMF), and showed that it has several useful properties: The spatial marginals are Gaussian Matérn fields; the family contains Markovian diffusion processes with clear physical interpretations; and we can control the smoothness in space and in time, the degree of non-separability, and interpret all the parameters. The family can also be extended to non-stationary models and be defined on curved manifolds.

The DEMF family contains several important subfamilies; 1) separable models, 2) Markov models, 3) partially separable models, 4) a fully non-separable subfamily of the Stein (2005) family, and 5) spatially non-stationary model dynamics. This provides a rich outset for studying the practical and methodological impacts these assumptions have.

An important special case in the DEMF family is the DEMF(1,2,1) model, which in two-dimensional space, is the closest stochastic process analogue to the diffusion equation (see (4)), and hence a natural default choice for spatio-temporal model components. The non-separable DEMF(1,2,1) model has the same smoothness in space and in time as the separable DEMF(1,0,2) model, which has a covariance function that is a Kronecker product of a Matérn covariance in space and an exponential covariance in time. Of particular interest is also the non-separable DEMF(2,2,0) model which can be viewed as an iterated diffusion model.

Although the proposed model family includes non-separable models, which in itself might be desirable from considerations about covariance properties, another view-point is that the non-separability here arises as a direct and natural consequence of the physics-inspired dynamical diffusion construction. Most importantly, the results shed light on which types of non-separability would occur naturally under certain assumptions on the spatio-temporal dynamics and properties of the driving noise process. Although there are strong arguments in the literature against using a separable model, the space of non-separable models is vastly larger than the space of separable models. Hence we need to consider which types of non-separable models are more, and which are less, appropriate than the separable alternatives. As illustrated by the practical example in Section 5.2, it is important to assess models in a context relevant to the intended use case. In particular, non-separability is unlikely to make a difference for space-time interpolation, as assessed by e.g. leave-one-out cross-validation, but can make a difference in full space-time forecasting settings.

It is natural to view the model class as an example of building models via building blocks with precision operator space-time separability. The most basic form of separability is *functional separability*, where a spatial and temporal processes are added or multiplied, which can be viewed as having $S+T$ degrees of freedom, where S and T are the spatial and temporal effective dimensions of the functions. The next form is *covariance separability*, where the model is formed from a sum of covariances (giving the same as functional separability) or a product of covariances, where the latter gives $S \cdot T$ degrees of freedom. These covariance product models are covariance separable but functionally non-separable. For precision models, plain products are equivalent to covariance separable models, but sums of precision products give covariance non-separability. In both the covariance and precision cases, non-stationarity in the spatial and temporal operators can be introduced, as long as the operator separability is kept. This distinguishes this type of non-separability from fully non-separable non-stationary models that cannot be written as precision sums and products. The key is to

retain commutativity between the spatial and temporal operators within each product: $(\boldsymbol{Q}_t \otimes \boldsymbol{I}_s)(\boldsymbol{I}_t \otimes \boldsymbol{Q}_s) = (\boldsymbol{I}_t \otimes \boldsymbol{Q}_s)(\boldsymbol{Q}_t \otimes \boldsymbol{I}_s) = \boldsymbol{Q}_t \otimes \boldsymbol{Q}_s$.

With the GMRF representation presented herein, the computational costs of the separable and non-separable models are similar, as the sparsity structure of posterior precisions, given irregularly spaced observations in generalised latent Gaussian models, is only marginally affected by the non-separability, and can even be more sparse in the non-separable cases; the separable precision neighbourhood structures are space-time prisms, whereas the non-separable neighbourhood structures are double-cones. Together with interpretable parameters, this makes the non-separable models as practically accessible as the separable models. In the supplementary materials we provide an implementation with examples in `R-INLA`.

In this paper we mainly focused on stationary fields, but also showed how very little in the theory and computational construction changes for models with curved manifolds or spatially non-stationary operators, as already discussed by Lindgren, Rue and Lindström (2011). Although the initial practical implementation only covers a subset of the general model class, we believe that the general results can and will be applied in more general contexts in the future.

Supplementary Material

R code for the examples: The examples were computed with the `INLAspacetime` package, using the `cgeneric` method from the `R-INLA` software for computationally efficiency, via the `inlabru` interface Bachl et al. (2019). See also Niekerk et al. (2021) for a similar example. Code for the figures and examples is available at https://github.com/finnlindgren/spacetime-paper-code, and the `INLAspacetime` R package (https://github.com/eliaskrainski/INLAspacetime) implements a subset of the models.

References

Bachl, F. E., Lindgren, F., Borchers, D. L. and Illian, J. B. (2019). inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10, 760-766.

Bakka, H., Rue, H., Fuglstad, G.-A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D. and Lindgren, F. (2018). Spatial modeling with R-INLA: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10, e1443.

Bakka, H., Vanhatalo, J., Illian, J. B., Simpson, D. and Rue, H. (2019). Non-stationary Gaussian models with physical barriers. *Spatial Statistics*, 29, 268-288.

Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K. E., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briët, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibul-

skis, R. E. and Gething, P. W. (2015). The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, 526, 207-211.

Bissiri, P. G., Holmes, C. C. and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 1103-1130.

Bolin, D. (2014). Spatial Matérn fields driven by non-Gaussian noise. *Scandinavian Journal of Statistics*, 41, 557-579.

Bolin, D. and Kirchner, K. (2020). The Rational SPDE Approach for Gaussian Random Fields With General Smoothness. *Journal of Computational and Graphical Statistics*, 29, 274-285. Publisher: Taylor & Francis.

Bolin, D., Simas, A. B. and Xiong, Z. (2024). Covariance-based rational approximations of fractional SPDEs for computationally efficient Bayesian inference. *Journal of Computational and Graphical Statistics*, 33, 64-74.

Bolin, D. and Wallin, J. (2023). Local scale invariance and robustness of proper scoring rules. *Statistical Science*, 38, 140-159.

Cameletti, M., F. Lindgren, D. Simpson, and H. Rue (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, 97, 109-131.

Clarotto, L., Allard, D., Romary, T. and Desassis, N. (2022). The SPDE approach for spatio-temporal datasets with advection and diffusion. *arXiv 2208.14015*.

Cramér, H. and Leadbetter, M. R. (1967). *Stationary and related stochastic processes: Sample function properties and their applications*. Wiley. Also available as a Dover reprint (2004).

Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94, 1330-1339.

Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ.

Da Prato, G. and J. Zabczyk (2014). *Stochastic equations in infinite dimensions*. Cambridge university press.

Erdélyi, A. (1953). Higher transcendental functions. *Higher Transcendental Functions, Edited by A. Erdelyi. Vol. I*, p. 59. McGraw-Hill.

Fonseca, T. C. and Steel, M. F. (2011). A general class of nonseparable space–time covariance models. *Environmetrics*, 22, 224-242.

Fuentes, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika*, 89, 197-210.

Fuentes, M., Chen, L. and Davis, J. M. (2008). A class of nonseparable and nonstationary spatial temporal covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, 19, 487-507.

Fuglstad, G.-A. and Castruccio, S. (2020). Compression of climate simulations with a nonstationary global spatiotemporal SPDE model. *Annals of Applied Statistics*, 14, 542-559.

Fuglstad, G. A., Simpson, D., Lindgren, F. and Rue, H. (2018). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 0, 1-8.

Gaedke-Merzhäuser, L., van Niekerk, J., Schenk, O. and Rue, H. (2022). Parallelized integrated nested laplace approximations for fast bayesian inference. *Statistics and Computing*, 33.

Gelfand, A., Diggle, P., Guttorp, P. and Fuentes, M. (2010). *Handbook of Spatial Statistics*. CRC Press.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97, 590-600.

Gneiting, T., Raftery, A., Westveld III, A. and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098-1118.

Guttorp, P. and Gneiting, T. (2006). Studies in the history of probability and statistics XLIX On the Matérn correlation family. *Biometrika*, 93, 989-995.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, 35, 403-410.

Held, L., Schrodle, B. and Rue, H. v. (2010). Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA. In *Statistical Modelling and Regression Structures*, pp. 111–131. Springer.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In C. W. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi (Eds.), *Quantitative Methods for Current Environmental Issues*, London, pp. 37–56. Springer London.

Horrell, M. T. and Stein, M. L. (2017). Half-spectral space-time covariance models. *Spatial Statistics*, 19, 90-100.

Jones, R. H. and Zhang, Y. (1997). Models for continuous stationary space-time processes. In T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren, and R. D. Wolfinger (Eds.), *Modelling Longitudinal and Spatially Correlated Data*, New York, NY, pp. 289–298. Springer New York.

Kelbert, M. Y., Leonenko, N. N. and Ruiz-Medina, M. D. (2005). Fractional random fields associated with stochastic fractional heat equations. *Advances in Applied Probability*, 37, 108-133.

Krainski, E. T. (2018). *Statistical Analysis of Space-time Data: New Models and Applications*. Ph. D. thesis, Norwegian University of Science and Technology.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilio, D., Simpson, D., Lindgren, F. and Rue, H. (2019). *Advanced Spatial Modeling with Stochastic Partial Differential Equations using R and INLA*. New York: Chapman and Hall/CRC. Github version www.r-inla.org/spde-book.

Lindgren, F., Bolin, D. and Rue, H. (2022). The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spatial Statistics*, 50, 100599.

Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation

approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 423-498.

Lindgren, G. (2012). *Stationary Stochastic Processes: Theory and Applications*. Chapman and Hall.

Liu, X., Guillas, S. and Lai, M.-J. (2016). Efficient Spatial Modeling Using the SPDE Approach With Bivariate Splines. *Journal of Computational and Graphical Statistics*, 25, 1176-1194.

Liu, X., Yeo, K. and Lu, S. (2022). Statistical Modeling for Spatio-Temporal Data From Stochastic Convection-Diffusion Processes. *Journal of the American Statistical Association*, 117, 1482-1499. Publisher: Taylor & Francis.

Matérn, B. (1960). Spatial variation - stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden från statens skogsforskningsinstitut*, 49.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E. and Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29, 897-910.

Moraga, P. (2019). *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. CRC Press.

Niekerk, J. V., Bakka, H., Rue, H. and Schenk, O. (2021). New Frontiers in Bayesian Modeling Using the INLA Package in R. *Journal of Statistical Software*, 100, 1-28.

Porcu, E., Furrer, R. and Nychka, D. (2021). 30 years of space-time covariance functions. *Wiley Interdisciplinary Reviews. Computational Statistics (WIREs)*, 13, e1512, 24.

Prévôt, C. and Röckner, M. (2007). *A concise course on stochastic partial differential equations*, Volume 1905. Springer.

Rodrigues, A. and Diggle, P. J. (2010). A class of convolution-based models for spatio-temporal processes with non-separable covariance structure. *Scandinavian Journal of Statistics*, 37, 553-567.

Roques, L., Allard, D. and Soubeyrand, S. (2022). Spatial statistics and stochastic partial differential equations: A mechanistic viewpoint. *Spatial Statistics*, 50, 100591.

Rozanov, J. A. (1977). Markov random fields and stochastic partial differential equations. *Mathematics of the USSR-Sbornik*, 32, 515-534.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 319-392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. and Lindgren, F. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, 4, 395-421.

Scheuerer, M. (2010). Regularity of the sample paths of a general second order random field. *Stochastic Processes and their Applications*, 120, 1879-1897.

Serra, L., Saez, M., Juan, P., Varga, D. and Mateu, J. (2014). A spatio-temporal Poisson hurdle point process to model wildfires. *Stochastic Environmental Research and Risk Assessment*, 28, 1671-1684.

Sigrist, F., Künsch, H. R. and Stahel, W. A. (2015). Stochastic partial differential equation based modelling of large space–time data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 3-33.

Simpson, D. P., Rue, H., Riebler, A., Martins, T. G. and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). *Statistical Science*, 32, 1-28.

Stein, M. L. (2005). Space–time covariance functions. *Journal of the American Statistical Association*, 100, 310-321.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Storvik, G., Frigessi, A. and Hirst, D. (2002). Stationary space-time gaussian fields and their time autoregressive representation. *Statistical Modelling*, 2, 139-161.

van Niekerk, J., Bakka, H., Rue, H. and Schenk, O. (2021). New frontiers in Bayesian modeling using the INLA package in R. *Journal of Statistical Software*, 100, 1-28.

van Niekerk, J., Krainksi, E., Rustand, D. and Rue, H. (2023). A new avenue for Bayesian inference with INLA. *Computational Statistics & Data Analysis*, 181, 107692.

van Niekerk, J. and Rue, H. (2024). Low-rank variational bayes correction to the Laplace method. *Journal of Machine Learning Research*.

Vergara, R. C., Allard, D. and Desassis, N. (2022). A general framework for SPDE-based stationary random fields. *Bernoulli*, 28, 1-32. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

Wahba, G. (1981). Spline Interpolation and Smoothing on the Sphere. *SIAM Journal on Scientific and Statistical Computing*, 2, 5-16. Publisher: Society for Industrial and Applied Mathematics.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41, 434-449.

Whittle, P. (1963). Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, 40, 974-994.

Whittle, P. (1986). *Systems in stochastic equilibrium*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Wikle, C. K. (2015). Modern perspectives on statistics for spatio-temporal data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7, 86-98.

Wood, A. W., Leung, L. R., Sridhar, V. and Lettenmaier, D. P. (2004). Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, 62, 189-216.

Yuan, Y., Bachl, F. E., Lindgren, F., Borchers, D. L., Illian, J. B., Buckland, S. T., Rue, H. and Gerrodette, T. (2017). Point process models for spatio-temporal distance sampling

data from a large-scale survey of blue whales. *The Annals of Applied Statistics*, 11, 2270-2297.

Zuur, A. F., Ieno, E. N. and Saveliev, A. A. (2017). Spatial, temporal and spatial-temporal ecological data analysis with R-INLA. *Highland Statistics Ltd*, 1.

Appendices

A. Almost sure sample path continuity

We start by rephrasing the main theorem of Section 9.3 of Cramér and Leadbetter (1967), and giving a formal definition of the smoothness index.

Definition A.1 (Cramér and Leadbetter, Section 2.5, generalised)**.** *A stochastic process $x(t)$ on some domain $\mathcal{D}$, is* equivalent *to another process $y(t)$ on $\mathcal{D}$, if for each fixed $t \in \mathcal{D}$, $x(t) = y(t)$, with probability one. This means that x differs from y on at most a set with measure zero, and that they have the same finite dimensional distributions.*

This technical definition allows us to view equivalent processes as an equivalence class that encapsulates some of the finer details of probabilistic measure theory for sample path continuity of stochastic processes.

Theorem A.1 (Cramér and Leadbetter, Section 9.3)**.** *Let $S^*(\omega)$ be the spectral measure of a stationary Gaussian process $x(t)$ on $t \in \mathbb{R}$, and let*

$$I_{a,b} = \int_0^\infty \omega^{2a} \left[\log(1+\omega)\right]^b \mathrm{d}S^*(\omega)$$

for $a, b \geq 0$. For spectral measures that admit a spectral density $S(\omega)$, replace $\mathrm{d}S^*(\omega)$ in $I_{a,b}$ with $S(\omega)\,\mathrm{d}\omega$.

1. *If $I_{a,b} < \infty$ for some $b > 3$ and some a in the range $[k, k+1)$ for some $k \in \mathbb{N}$, then $x(t)$ is equivalent to a process $y(t)$ that has a continuous sample derivative of order k, with probability one.*

2. *If $I_{a,1} < \infty$ for some a in the range $(k, k+1]$ for some $k \in \mathbb{N}$, then $x(t)$ is equivalent to a process $y(t)$ whose sample derivative of order k is Hölder continuous with exponent $a - k \in (0, 1]$, with probability one.*

For the case $k = 0$, the sample derivative of order zero refers to the sample path of the process itself.

Proof. The results follow directly from the main theorem of Section 9.3 of Cramér and Leadbetter (1967). ■

Results from Scheuerer (2010) show that under a similar condition for d-dimensional domains,

$$\int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|^{2a} [\log(1+\|\boldsymbol{\omega}\|)]^b \, \mathrm{d}S^*(\boldsymbol{\omega}) < \infty,$$

for all $a < \nu$ and some $b > 1$, the sample paths on $\mathbb{R}^d$ belong to any Sobolev space $W^{a,2}$ of order $a < \nu$, on any bounded subdomain, with probability one. For isotropic spectra, this translates to $I_{a,b} < \infty$ for all $a < \nu$ and some $b > 1$, when applied to the one-dimensional marginal spectra.

The integral criteria above motivate the following characterisation of the *smoothness index* ν, in particular when applied to models with power law spectral density tails.

Definition A.2. *The smoothness index ν of a stationary Gaussian process $x(t)$, $t \in \mathbb{R}$, is $\nu = \sup_a\{a; I_{a,1} < \infty\}$, where $I_{a,1}$ is defined as in Theorem A.1.*

B. Numerical evaluation of covariances

When spatio-temporal spectral density is available in closed format on $\mathbb{R}^d \times \mathbb{R}$, the covariance function can be obtained to close numerical accuracy using fast Fourier transformation (FFT). In order to reduce the memory requirements for isotropic models on high-dimensional spatial domains, the marginal space-time spectrum along a single spatial dimension can be evaluated first. For general models, evaluating spatial FFT transformations for each time lag further reduces the memory footprint if only some of the covariances are stored.

The idea is construct the folded spectrum resulting from spatial/temporal discretisation, and then discretise it onto a finite regular lattice. The resulting integral approximations can be evaluated with standard FFT implementations, and the numerical approximation error in the covariance evaluation is determined by the frequency resolution and smoothness of the spectral density. The brief theory behind the construction presented below is based on Lindgren (2012).

B.1. Spectral folding

The exact spectral representation of the covariance evaluated on a discrete infinite lattice can be derived from the continuous domain representation. For simplicity, assume the same lattice spacing h in each direction. A stationary covariance function $R(\boldsymbol{s})$ evaluated

at lattice points $\boldsymbol{j}h$, $\boldsymbol{j} \in \mathbb{Z}^d$ is given by

$$\begin{aligned} R(\boldsymbol{s}) &= \int_{\mathbb{R}^d} \exp(i\boldsymbol{\omega}\cdot\boldsymbol{s}) S(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega}, \\ R(\boldsymbol{j}h) &= \int_{\mathbb{R}^d} \exp(i\boldsymbol{\omega}\cdot\boldsymbol{j}h) S(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega} \\ &= \int_{[-\pi/h,\pi/h)^d} \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \exp(i(\boldsymbol{\omega}+2\pi\boldsymbol{k}/h)\cdot\boldsymbol{j}h) S(\boldsymbol{\omega}+2\pi\boldsymbol{k}/h)\,\mathrm{d}\boldsymbol{\omega} \\ &= \int_{[-\pi/h,\pi/h)^d} \exp(i\boldsymbol{\omega}\cdot\boldsymbol{j}h) \widetilde{S}(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega}, \end{aligned} \tag{27}$$

where

$$\widetilde{S}(\boldsymbol{\omega}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} S(\boldsymbol{\omega}+2\pi\boldsymbol{k}/h), \quad \boldsymbol{\omega} \in [-\pi/h,\pi/h)^d.$$

If instead the spatial discretisation should be interpreted as the *cell averages* (which is the more usual case for PDE discretisations and e.g. satellite data, rather than pointwise values), the spectrum is altered by a multiplicative frequency filter with a squared sinc function:

$$\widetilde{S}(\boldsymbol{\omega}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} S(\boldsymbol{\omega}+2\pi\boldsymbol{k}/h) \prod_{l=1}^{d} \left\{ \frac{\sin[(\omega_l+2\pi k_l/h)/2]}{(\omega_l+2\pi k_l/h)/2} \right\}^2, \quad \boldsymbol{\omega} \in [-\pi/h,\pi/h)^d.$$

B.2. Discrete Fourier transformation

To approximate the integral in (27) with FFT, choose a positive integer M . This gives a numerical integration approximation

$$\widehat{R}(\boldsymbol{j}h) = \left(\frac{\pi}{hM}\right)^d \sum_{\boldsymbol{k}\in[-M,M)^d} \exp\left(i\boldsymbol{k}\cdot\boldsymbol{j}\frac{2\pi}{2M}\right) \widetilde{S}\left(\boldsymbol{k}\frac{\pi}{hM}\right), \quad \boldsymbol{j} \in [-M,M)^d, \tag{28}$$

which is of the form that can be evaluated using FFT.

B.2.1. Sampling

With the above theory, sampling from the model can be expressed as an integral with respect to continuous domain complex valued white noise process, $\mathrm{d}Z(\boldsymbol{\omega})$, with conjugate symmetry:

$$x(\boldsymbol{j}h) = \int_{[-\pi/h,\pi/h)^d} \exp(i\boldsymbol{\omega}\cdot\boldsymbol{j}h) \widetilde{S}(\boldsymbol{\omega})^{1/2}\,\mathrm{d}Z(\boldsymbol{\omega}), \quad \boldsymbol{j} \in \mathbb{Z}^d,$$

where $\overline{\mathrm{d}Z(-\boldsymbol{\omega})} = \mathrm{d}Z(\boldsymbol{\omega})$, $\mathsf{Cov}(\mathrm{d}Z(\boldsymbol{\omega}),\mathrm{d}Z(\boldsymbol{\omega}')) = \delta(\boldsymbol{\omega}-\boldsymbol{\omega}')\,\mathrm{d}\boldsymbol{\omega}$. This can be discretised with a lattice of frequencies in much the same way as for computing the covariance function, with noise variances equal to the cell area/volume $\left(\frac{\pi}{hM}\right)^d$ of each frequency lattice point. When the outer pairwise opposing cells are discretised, the combined complex noise contributions are real, and should be assigned to the $-M$ indices, which ensures that the resulting field has no non-zero imaginary components.

C. Spherical harmonics

C.1. Definition and standard properties

In $\mathbb{R}^2$, the harmonic functions, sine and cosine, play an important role as basis functions in spectral representations of functions and random fields. On the sphere, this role is instead taken by the *spherical harmonics*. This section presents the basic results needed for spectral representation theory for stationary processes on the sphere.

Definition C.1. *The* spherical harmonic $Y_{k,m}(\boldsymbol{u})$, $\boldsymbol{u} = \left[u_1, u_2, u_3\right]^\top \in \mathbb{S}^2 \subset \mathbb{R}^3$, *of* order $k = 0, 1, 2, \ldots$ *and* mode $m = -k, \ldots, k$ *is defined by*

$$Y_{k,m}(\boldsymbol{u}) = \sqrt{(2k+1)\cdot\frac{(k-|m|)!}{(k+|m|)!}}\cdot\begin{cases}\sqrt{2}\sin(m\phi)P_{k,-m}(\cos\theta) & -k \le m < 0,\\ P_{k,0}(\cos\theta) & m = 0,\\ \sqrt{2}\cos(m\phi)P_{k,m}(\cos\theta) & 0 < m \le k,\end{cases}$$

where ϕ *is the longitude and* $\theta = \arccos(u_3)$ *is the colatitude, and* $P_{k,|m|}(u_3)$ *are* associated Legendre functions *($P_{k,0}(u_3)$ are Legendre polynomials). Note that* $\sin\phi = u_2/\sqrt{u_1^2+u_2^2}$, $\cos\phi = u_1/\sqrt{u_1^2+u_2^2}$, *and* $\cos\theta = u_3$.

Standard property results for spherical harmonics, following Wahba (1981), building the basis of spherical Fourier theory:

1. The spherical harmonics form an orthogonal basis for functions on the unit sphere, $\mathbb{S}^2$:

$$\langle Y_{k,m}, Y_{k',m'}\rangle_{\mathbb{S}^2} = \begin{cases}4\pi, & k' = k, m' = m,\\ 0, & \text{otherwise.}\end{cases}$$

2. The addition formula for spherical harmonics is

$$\sum_{m=-k}^{k} Y_{k,m}(\boldsymbol{u})Y_{k,m}(\boldsymbol{v}) = (2k+1)P_{k,0}(\boldsymbol{u}^\top\boldsymbol{v}).$$

3. The spherical harmonics are eigenfunctions to the Laplacian on $\mathbb{S}^2$,

$$\Delta Y_{k,m}(\boldsymbol{u}) = -k(k+1)Y_{k,m}(\boldsymbol{u}).$$

4. Let $\phi(\boldsymbol{u})$ be a square-integrable function on $\mathbb{S}^2$. Then $\phi(\boldsymbol{u})$ has series expansion

$$\phi(\boldsymbol{u}) = (\mathcal{F}^{-1}\widehat{\phi})(\boldsymbol{u}) = \sum_{k=0}^{\infty}\sum_{m=-k}^{k}\widehat{\phi}(k,m)Y_{k,m}(\boldsymbol{u}),$$

with *Fourier Bessel* coefficients $\widehat{\phi}(k,m) = (\mathcal{F}\phi)(k,m) = \frac{1}{4\pi}\langle\phi(\boldsymbol{u}), Y_{k,m}(\boldsymbol{u})\rangle_{\mathbb{S}^2(\mathrm{d}\boldsymbol{u})}$. Also, $\langle\phi, 1\rangle_{\mathbb{S}^2} = 4\pi\widehat{\phi}(0,0)$ and $\langle\phi,\phi\rangle_{\mathbb{S}^2} = 4\pi\sum_{k,m}\widehat{\phi}(k,m)^2$.

C.2. Spherical variance approximation

Define

$$F_{a,b} = \sum_{k=a}^{b} \frac{2k+1}{4\pi[\gamma_s^2 + k(k+1)]^\alpha},$$

so that $F_{0,\infty}$ gives the variance in (23). With

$$\begin{aligned} I_{a,b} &= \int_a^b \frac{2x+1}{4\pi[\gamma_s^2 + x(x+1)]^\alpha}\,\mathrm{d}x \\ &= \frac{1}{4\pi(\alpha-1)}\left(\frac{1}{[\gamma_s^2 + a(a+1)]^{\alpha-1}} - \frac{1}{[\gamma_s^2 + b(b+1)]^{\alpha-1}}\right), \end{aligned}$$

choose K so that the terms in the sum (23) are decreasing for $k \geq K$. This holds for any $K \geq K_0$, where $K_0 = 0$ if $\gamma_s \leq 1/2$, and $K_0 = \left\lceil \sqrt{\frac{\gamma_s^2 - 1/4}{2\alpha - 1}} - \frac{1}{2} \right\rceil$ for $\gamma_s > 1/2$. Then the full sum $F_{0,\infty}$ can be bounded by a partial sum $F_{0,K}$ and tail integrals:

$$F_{0,K} + I_{K+1,\infty} \leq F_{0,\infty} \leq F_{0,K} + I_{K,\infty}.$$

Tighter bounds can in principle be obtained for the approximation $F_{0,\infty} \approx F_{0,K} + I_{K+1/2,\infty}$. Let f_x denote the integrand for $I_{a,b}$. Then a second order Taylor expansion around each $x = k$ gives the error bound

$$F_{0,K} + I_{K+1/2,\infty} - F_{0,\infty} \;=\; I_{K+1/2,\infty} - F_{K+1,\infty} \;\leq \frac{1}{24} \sum_{k=K+1}^{\infty} \sup_{x \in (k-1/2, k+1/2)} f_x'' \;.$$

It may be possible to construct a bound for this series using another integral bound, but the practical utility of doing so is unclear.

D. Collected proofs

D.1. Proof of Proposition 3.1

The covariance function for spatial lag $\boldsymbol{s} = \boldsymbol{s}_2 - \boldsymbol{s}_1$ and temporal lag t can be written as a nested integral,

$$\begin{aligned} \mathrm{cov}[u(\mathbf{0},0), u(\boldsymbol{s},t)] &= \int_{\mathbb{R}^d} \int_{\mathbb{R}} \exp[i(\boldsymbol{s}\cdot\boldsymbol{\omega}_s + t\omega_t)] S_u(\boldsymbol{\omega}_s, \omega_t)\,\mathrm{d}\omega_t\,\mathrm{d}\boldsymbol{\omega}_s \\ &= \int_{\mathbb{R}^d} \exp(i\boldsymbol{s}\cdot\boldsymbol{\omega}_s) \left\{ \int_{\mathbb{R}} \exp(it\omega_t) S_u(\boldsymbol{\omega}_s, \omega_t)\,\mathrm{d}\omega_t \right\} \mathrm{d}\boldsymbol{\omega}_s. \\ &= \int_{\mathbb{R}^d} \exp(i\boldsymbol{s}\cdot\boldsymbol{\omega}_s) S_u(\boldsymbol{\omega}_s; t)\,\mathrm{d}\boldsymbol{\omega}_s, \end{aligned}$$

where the inner integral $S_u(\boldsymbol{\omega}_s; t)$ is the marginal spatial cross-spectrum for time lag t.

Let $\lambda = \gamma_s^2 + \|\boldsymbol{\omega}_s\|^2$ and $\kappa^2 = \lambda^{\alpha_s}/\gamma_t^2$. Then, integrating over ω_t, we get

$$\begin{aligned}
S_u(\boldsymbol{\omega}_s;t) &= \frac{1}{(2\pi)^d \gamma_e^2 \lambda^{\alpha_e} \gamma_t^{2\alpha_t}} \int_{\mathbb{R}} \frac{\mathrm{e}^{it\omega_t}}{2\pi(\omega_t^2 + \lambda^{\alpha_s}/\gamma_t^2)^{\alpha_t}} \,\mathrm{d}\omega_t \\
&= \frac{1}{(2\pi)^d \gamma_e^2 \lambda^{\alpha_e} \gamma_t^{2\alpha_t}} \frac{C_{\mathbb{R},\alpha_t}}{\kappa^{2(\alpha_t-1/2)}} R^M_{\alpha_t-1/2}(\kappa t) \\
&= \frac{C_{\mathbb{R},\alpha_t}}{\gamma_e^2 \gamma_t} \frac{1}{(2\pi)^d (\gamma_s^2 + \|\boldsymbol{\omega}_s\|^2)^{\alpha}} R^M_{\alpha_t-1/2}\left\{ t\sqrt{\gamma_s^2 + \|\boldsymbol{\omega}_s\|^2}/\gamma_t \right\},
\end{aligned}$$

where $R^M_\nu(t)$ is the standard Matérn correlation with smoothness ν, defined in (16), and the additional scaling was given in Lindgren et al. (2011). For $t = 0$, the temporal contribution factor is 1, and we recognize the resulting expression as the spectral density corresponding to a spatial Matérn covariance function with range parameter γ_s and smoothness parameter $\nu_s = \alpha - d/2$, and marginal variance equal to the sought value σ^2 in the proposition. We then also know that the marginal spectrum for $t = 0$ in any single spatial dimension is proportional to $(\gamma_s^2 + \omega^2)^{-\nu_s+1/2}$, which shows that the conditions on a in Theorem A.1 are fulfilled if and only if $a < \nu_s$, so ν_s is the smoothness index.

D.2. Proof of Proposition 3.2

Let ν_t be the smoothness index for the marginal temporal process $u(\boldsymbol{s},t)$. We need to identify for which values of a the integral $I_{a,1} = \int_0^\infty \omega_t^{2a} \log(1+\omega_t) S_u(\omega_t)\,\mathrm{d}\omega_t$ in Theorem A.1 is finite. We start by integrating out the spatial spectral dimensions and reparameterising the resulting integral:

$$\begin{aligned}
S_u(\omega_t) &\propto \int_{\mathbb{R}^d} [\gamma_t^2\omega_t^2 + (\gamma_s^2 + \|\boldsymbol{\omega}_s\|^2)^{\alpha_s}]^{-\alpha_t} (\gamma_s^2 + \|\boldsymbol{\omega}_s\|^2)^{-\alpha_e} \,\mathrm{d}\boldsymbol{\omega}_s \\
&\propto \int_0^\infty r^{d-1} [\gamma_t^2\omega_t^2 + (\gamma_s^2 + r^2)^{\alpha_s}]^{-\alpha_t} (\gamma_s^2 + r^2)^{-\alpha_e} \,\mathrm{d}r \\
&\propto \int_0^\infty v^{(d-2)/2} (1+v)^{-\alpha_e} (\widetilde{\omega}_t^2 + (1+v)^{\alpha_s})^{-\alpha_t} \,\mathrm{d}v
\end{aligned} \tag{29}$$

where we in the second step changed to polar coordinates and in the third set $v = r^2/\gamma_s^2$ and $\widetilde{\omega}_t = \omega_t \gamma_t/\gamma_s^{\alpha_s}$. The integral (29) is finite for all $\widetilde{\omega}_t$ when $\alpha_e + \alpha_s\alpha_t > d/2$. Assuming $a < \nu_t$, we can then write the integral in the smoothness criterion as

$$\begin{aligned}
I_{a,1} &= \int_0^\infty \omega_t^{2a} \log(1+\omega_t) S_u(\omega_t)\,\mathrm{d}\omega_t \\
&= C_0 \int_0^\infty \widetilde{\omega}_t^{2a} \log\left(1 + \frac{\widetilde{\omega}_t \gamma_s^{\alpha_s}}{\gamma_t}\right) \int_0^\infty v^{(d-2)/2} (1+v)^{-\alpha_e} (\widetilde{\omega}_t^2 + (1+v)^{\alpha_s})^{-\alpha_t} \,\mathrm{d}v\,\mathrm{d}\widetilde{\omega}_t
\end{aligned}$$

for some constant C_0. Let $\varepsilon > 0$ such that $a + \varepsilon < \nu_t$. Then $\log\left(1 + \frac{\widetilde{\omega}_t \gamma_s^{\alpha_s}}{\gamma_t}\right) \le C_\varepsilon \widetilde{\omega}_t^{2\varepsilon}$ for all $\widetilde{\omega}_t > 0$ for some $C_\varepsilon > 0$. We can then bound $I_{a,1}$ and change the order of integration

since the integrands are positive:

$$I_{a,1} \leq C_0 C_\varepsilon \int_0^\infty \widetilde{\omega}_t^{2(a+\varepsilon)} \int_0^\infty v^{(d-2)/2}(1+v)^{-\alpha_e}(\widetilde{\omega}_t^2 + (1+v)^{\alpha_s})^{-\alpha_t}\,\mathrm{d}v\,\mathrm{d}\widetilde{\omega}_t$$
$$= C_0 C_\varepsilon \int_0^\infty v^{(d-2)/2}(1+v)^{-\alpha_e} \int_0^\infty \frac{\widetilde{\omega}_t^{2(a+\varepsilon)}}{(\widetilde{\omega}_t^2 + (1+v)^{\alpha_s})^{\alpha_t}}\,\mathrm{d}\widetilde{\omega}_t\,\mathrm{d}v.$$

The change of variables $w = \frac{\widetilde{\omega}_t}{(1+v)^{\alpha_s/2}}$ in the inner integral gives

$$I_{a,1} \leq C_0 C_\varepsilon \int_0^\infty v^{(d-2)/2}(1+v)^{-\alpha_e} \int_0^\infty \frac{w^{2(a+\varepsilon)}(1+v)^{(a+\varepsilon-\alpha_t)\alpha_s}}{(w^2+1)^{\alpha_t}}(1+v)^{\alpha_s/2}\,\mathrm{d}w\,\mathrm{d}v$$
$$= C_0 C_\varepsilon \int_0^\infty v^{(d-2)/2}(1+v)^{-\alpha_e-\alpha_s(\alpha_t-a-\varepsilon-1/2)} \int_0^\infty \frac{w^{2(a+\varepsilon)}}{(w^2+1)^{\alpha_t}}\,\mathrm{d}w\,\mathrm{d}v.$$

In this expression, the inner integral is a finite constant, C_w, when $2\alpha_t - 2a - 2\varepsilon > 1$, i.e., when $a + \varepsilon < \alpha_t - 1/2$. Since ε can be chosen arbitrarily small, we can make C_w finite for all $a < \alpha_t - 1/2$. The remaining integral has an integrable singularity at $v = 0$ for $d = 1$, and the integral is finite when $\alpha_e + \alpha_s(\alpha_t - a - \varepsilon - 1/2) - (d-2)/2 > 1$. Solving for a and again recognising that ε can be chosen arbitrarily small, we have now shown that $I_{a,1} < \infty$ when both $a < \alpha_t - 1/2$ and $a < \frac{\alpha_e + (\alpha_t - 1/2) - d/2}{\alpha_s} = \frac{\nu_s}{\alpha_s}$ hold. Therefore the temporal smoothness is given by $\nu_t = \min(\alpha_t - 1/2, \frac{\nu_s}{\alpha_s})$.

We now turn to the special case $d = 2$, where we can derive an explicit expression for the spectral density. Let $B(x,y)$ be the beta function,

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t.$$

Making the change of variables $1 + x = (1+v)^{\alpha_s}$ in (29) the marginal temporal spectrum becomes

$$S_u(\omega_t) \propto \int_0^\infty (1+x)^{-\frac{\alpha_e-1}{\alpha_s}-1}(\widetilde{\omega}_t^2 + 1 + x)^{-\alpha_t}\,\mathrm{d}x \qquad \text{[formula 3.197.9 in G\&R (p317)]}$$
$$\propto B\left(\frac{\alpha_e-1}{\alpha_s} + \alpha_t, 1\right) {}_2F_1\left(\alpha_t, \frac{\alpha_e-1}{\alpha_s} + \alpha_t, \frac{\alpha_e-1}{\alpha_s} + \alpha_t + 1; -\widetilde{\omega}_t^2\right),$$

because $\frac{\alpha_e-1}{\alpha_s} + \alpha_t = \frac{\nu_s}{\alpha_s} + \frac{1}{2} > 0$. Finally we verify that this spectrum yields the smoothness parameter implied by the general dimension result. Assuming that $a - b$ is not an integer, the hypergeometric function ${}_2F_1(a,b;c,z)$ for large values of z behaves like

$${}_2F_1(a,b,c,z) \sim c_1 z^{-a} + c_2 z^{-b} + \mathcal{O}(z^{-a-1}) + \mathcal{O}(z^{-b-1})$$

as $z \to \infty$. If $a - b$ is an integer we have to multiply z^{-a} or z^{-b} with $\log(z)$ (Erdélyi (1953) volume 1, section 2.3.2, page 76). This extra logarithmic factor will not make a difference for the final smoothness. Thus, we may write

$$S_t(\omega_t) = \mathcal{O}(\omega_t^{-2\alpha_t}) + \mathcal{O}\left(\omega_t^{-2(\frac{\alpha_e-1}{\alpha_s}+\alpha_t)}\right) = \mathcal{O}\left(\omega_t^{-2(\alpha_t + \frac{1}{\alpha_s}\min(0,\alpha_e-1))}\right)$$

for large ω_t. This decay rate is such that the conditions in Theorem A.1 are if and only if $a < \nu_t$ with

$$\nu_t = \frac{2(\alpha_t + \frac{1}{\alpha_s}\min(0, \alpha_e - 1)) - 1}{2} = \alpha_t + \frac{1}{\alpha_s}\min(0, \alpha_e - 1) - \frac{1}{2} = \min\left[\alpha_t - \frac{1}{2}, \frac{\nu_s}{\alpha_s}\right],$$

which completes the proof.

D.3. Proof of Theorem 4.1

Define the eigenvector matrix $\boldsymbol{V}$ and the eigenvalue (diagonal) matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{n_s})$ solving the generalised eigenvalue problem $\boldsymbol{K}_1\boldsymbol{V} = \boldsymbol{C}\boldsymbol{V}\boldsymbol{\Lambda}$. Since $\boldsymbol{K}_1$ and $\boldsymbol{C}$ are symmetric and $\boldsymbol{K}_1$ is positive definite, the eigenvectors can be chosen so that $\boldsymbol{V}^\top\boldsymbol{C}\boldsymbol{V} = \boldsymbol{I}$. For general $a = 0, 1, 2, \ldots$, $\boldsymbol{K}_{a+1} = \boldsymbol{K}_a\boldsymbol{C}^{-1}\boldsymbol{K}_1$, so that $\boldsymbol{K}_{a+1}\boldsymbol{V} = \boldsymbol{K}_a\boldsymbol{V}\boldsymbol{\Lambda}$. Recursion shows that $\boldsymbol{K}_a\boldsymbol{V} = \boldsymbol{C}\boldsymbol{V}\boldsymbol{\Lambda}^a$, which also holds for general $a \geq 0$, and $\boldsymbol{V}^\top\boldsymbol{K}_a\boldsymbol{V} = \boldsymbol{\Lambda}^a$.

For $\alpha_t = 1$, the temporal evolution of the spatial Hilbert space discretisation of (25) is determined by

$$\left(\gamma_t\boldsymbol{C}\frac{\partial}{\partial t} + \boldsymbol{K}_{\alpha_s/2}\right)\boldsymbol{u}(t) = \boldsymbol{C}\,\mathrm{d}\mathcal{E}_{\gamma_e^2\boldsymbol{K}_{\alpha_e}}(t), \quad t \in \mathbb{R}.$$

A multivariate change of variables $\boldsymbol{u}(t) = \boldsymbol{V}\boldsymbol{z}(t)$ and multiplication by $\boldsymbol{V}^\top$ on both sides gives

$$\left(\gamma_t\boldsymbol{I}\frac{\partial}{\partial t} + \boldsymbol{\Lambda}^{\alpha_s/2}\right)\boldsymbol{z}(t) = \boldsymbol{V}^\top\boldsymbol{C}\,\mathrm{d}\mathcal{E}_{\gamma_e^2\boldsymbol{K}_{\alpha_e}}(t) = \mathrm{d}\mathcal{E}_{\gamma_e^2\boldsymbol{\Lambda}^{\alpha_e}}(t),$$

where the precision of the driving noise process follows from

$$\begin{aligned}\gamma_e^2\left(\boldsymbol{V}^\top\boldsymbol{C}\boldsymbol{K}_{\alpha_e}^{-1}\boldsymbol{C}\boldsymbol{V}\right)^{-1} &= \gamma_e^2\boldsymbol{V}^{-1}\boldsymbol{C}^{-1}\boldsymbol{K}_{\alpha_e}\boldsymbol{C}^{-1}\boldsymbol{V}^{-\top} \\ &= \gamma_e^2\boldsymbol{V}^\top\boldsymbol{K}_{\alpha_e}\boldsymbol{V} = \gamma_e^2\boldsymbol{\Lambda}^{\alpha_e}.\end{aligned}$$

For $\alpha_t = 2$, the same technique yields

$$\left(-\gamma_t^2\boldsymbol{C}\frac{\partial^2}{\partial t^2} + \boldsymbol{K}_{\alpha_s}\right)\boldsymbol{u}(t) = \boldsymbol{C}\,\mathrm{d}\mathcal{E}_{\gamma_e^2\boldsymbol{K}_{\alpha_e}}(t)$$

and

$$\left(-\gamma_t^2\boldsymbol{I}\frac{\partial^2}{\partial t^2} + \boldsymbol{\Lambda}^{\alpha_s}\right)\boldsymbol{z}(t) = \mathrm{d}\mathcal{E}_{\gamma_e^2\boldsymbol{\Lambda}^{\alpha_e}}(t).$$

Using the solutions for $\alpha_t = 1$ and 2 as the driving noise processes on the right hand side, the recursive construction technique from Lindgren et al. (2011) gives the general

spatial discretisations

$$\left(-\gamma_t^2 \boldsymbol{C}\frac{\partial^2}{\partial t^2} + \boldsymbol{K}_{\alpha_s}\right)^{\alpha_t/2} \boldsymbol{u}(t) = \boldsymbol{C}\,\mathrm{d}\mathcal{E}_{\gamma_e^2 \boldsymbol{K}_{\alpha_e}}(t),$$
$$\left(-\gamma_t^2 \boldsymbol{I}\frac{\partial^2}{\partial t^2} + \boldsymbol{\Lambda}^{\alpha_s}\right)^{\alpha_t/2} \boldsymbol{z}(t) = \mathrm{d}\mathcal{E}_{\gamma_e^2 \boldsymbol{\Lambda}^{\alpha_e}}(t),$$

for any $\alpha_t = 1, 2, \ldots$. Since the evolution of $\boldsymbol{z}(t)$ is independent between the vector components, we get

$$\left(-\gamma_t^2 \frac{\partial^2}{\partial t^2} + \lambda_i^{\alpha_s}\right)^{\alpha_t/2} z_i(t) = \frac{1}{\gamma_e \lambda_i^{\alpha_e/2}} \mathcal{W}_i(t), \quad \text{for } i = 1, \ldots, n_s,$$

where λ_i is the i:th generalised eigenvalue of $\boldsymbol{K}_1$, and $\mathcal{W}_i(\cdot)$ are white noise processes, independent across all i. Rearranging factors, we get

$$\gamma_e \lambda_i^{\alpha_e/2} \gamma_t^{\alpha_t} \left(-\frac{\partial^2}{\partial t^2} + \gamma_t^{-2} \lambda_i^{\alpha_s}\right)^{\alpha_t/2} z_i(t) = \mathcal{W}_i(t), \quad \text{for } i = 1, \ldots, n_s.$$

Applying the temporal condition of the theorem with $b_i = \gamma_e^2 \lambda_i^{\alpha_e} \gamma_t^{2\alpha_t}$ and $\kappa_i = \lambda_i^{\alpha_s/2}/\gamma_t$ then gives a the temporal discretisation precision for each $z_i(t)$ as

$$\boldsymbol{Q}_{z_i} = \sum_{k=0}^{2\alpha_t} b_i \kappa_i^{2\alpha_t - k} \boldsymbol{J}_{\alpha_t, k/2}.$$

Collecting the processes gives the joint precision as

$$\boldsymbol{Q}_{\boldsymbol{z}} = \sum_{k=0}^{2\alpha_t} \boldsymbol{J}_{\alpha_t, k/2} \otimes \mathrm{diag}(b_i \kappa_i^{2\alpha_t - k}) = \gamma_e^2 \sum_{k=0}^{2\alpha_t} \gamma_t^k \boldsymbol{J}_{\alpha_t, k/2} \otimes \boldsymbol{\Lambda}^{\alpha_e + (2\alpha_t - k)\alpha_s/2}.$$

The joint discretisation vector in the original parameterisation is given by $\boldsymbol{u} = (\boldsymbol{I} \otimes \boldsymbol{V})\boldsymbol{z}$, with covariance $\boldsymbol{Q}_{\boldsymbol{u}}^{-1} = (\boldsymbol{I} \otimes \boldsymbol{V})\boldsymbol{Q}_{\boldsymbol{z}}^{-1}(\boldsymbol{I} \otimes \boldsymbol{V}^\top)$. We note that $\boldsymbol{V}^{-\top} \boldsymbol{\Lambda}^a \boldsymbol{V}^{-1} = \boldsymbol{K}_a$, so that the joint precision matrix becomes

$$\boldsymbol{Q}_{\boldsymbol{u}} = (\boldsymbol{I} \otimes \boldsymbol{V}^{-\top})\boldsymbol{Q}_{\boldsymbol{z}}(\boldsymbol{I} \otimes \boldsymbol{V}^{-1}) = \gamma_e^2 \sum_{k=0}^{2\alpha_t} \gamma_t^k \boldsymbol{J}_{\alpha_t, k/2} \otimes \boldsymbol{K}_{\alpha_e + (\alpha_t - k/2)\alpha_s},$$

which completes the proof.

E. Temporal GMRF representation with stationary boundary conditions

We present precision matrices for stationary AR(2) (autogregressive order 2) processes, and then show how this can be used to construct stationary boundary conditions for GMRF representations of 1st and second order Whittle-Matérn type stochastic differential equations.

Lemma E.1. *Let u_k be a stationary AR(2) process with evolution*

$$a_0 u_k + a_1 u_{t-k} + a_2 u_{k-2} = e_k,$$

with $a_0 > 0$ and e_k independent, $e_k \sim N(0,1)$. Then, the precision matrix $\boldsymbol{Q}$ for $(u_1, \ldots, u_n)$ is quint-diagonal, and, except for the upper left and lower right 2×2 corners, $\boldsymbol{Q}$ has diagonal elements $q_0 = a_0^2 + a_1^2 + a_2^2$ and off-diagonal elements $q_1 = a_1(a_0 + a_2)$ and $q_2 = a_0 a_2$. Further, the corner elements are given by

$$\begin{aligned} Q_{0,0} &= Q_{n,n} = a_0^2, & Q_{1,1} &= Q_{n-1,n-1} = a_0^2 + a_1^2, \\ Q_{0,1} &= Q_{n,n-1} = a_1 a_0, & Q_{1,0} &= Q_{n-1,n} = a_1 a_0. \end{aligned}$$

Conversely, if the inner elements q_0, q_1, and q_2 are known, the a_0, a_1, and a_2 values can be recovered, and hence the corner elements be constructed: Define the constants

$$b_+ = \sqrt{q_0 + 2q_1 + 2q_2}, \quad b_- = \sqrt{q_0 - 2q_1 + 2q_2}, \quad b_s = \frac{b_+ + b_-}{2}.$$

Then,

$$a_0 = \frac{1}{2}\left(b_s + \sqrt{b_s^2 - 4q_2}\right), \quad a_1 = \frac{b_+ - b_-}{2}, \quad a_2 = \frac{1}{2}\left(b_s - \sqrt{b_s^2 - 4q_2}\right).$$

Proof. Follows by direct computation. ∎

Let $\Phi_t = \{\phi_1(t), ..., \phi_{N_t}(t)\}$ be a set of piecewise linear basis functions in time, on a regular grid, and consider precision matrices on the coefficients for a linear combination of these basis functions. We want to obtain a GMRF representation of a stationary process Ornstein-Uhlenbeck process $z(t)$, such that

$$\kappa z(t) + \frac{d}{dt} z(t) = b^{-1/2} \varepsilon(t), \quad t \in \mathbb{R} \tag{30}$$

where ε is white noise. However, we can instead use the equivalent stochastic process model

$$\left(\kappa^2 - \frac{d^2}{dt^2}\right)^{1/2} z(t) = b^{-1/2} \varepsilon(t), \quad t \in \mathbb{R}. \tag{31}$$

Under stationarity, these two models are equivalent in the sense that they have the same covariance function. Let $\boldsymbol{M}_0 = (\langle \phi_i, \phi_j \rangle)_{i,j}$, $\boldsymbol{M}_2 = (\langle \nabla \phi_i, \nabla \phi_j \rangle)_{i,j}$. Assuming Neumann boundary conditions on a finite interval, and (31), the precision matrix is $\boldsymbol{Q} = b(\kappa^2 \boldsymbol{M}_0 + \boldsymbol{M}_2)$, see Lindgren et al. (2011, Sec 2.3). This matrix does not represent a stationary process on the finite interval. However, it is quint-diagonal, and can be corrected to give a stationary GMRF by adding $b\kappa\sqrt{1 + h^2\kappa^2/4} \approx b\kappa$, to the first and the last entries of

the matrix $\boldsymbol{Q}$, per the previous lemma. Here, h is the step-size in the mesh, and we assume that $h\kappa$ is small. Let $\boldsymbol{M}_1$ be a matrix of zeroes, except the first and last elements which are $1/2$. We then have a stationary GMRF representation of the process (31) with precision matrix

$$b(\kappa^2\boldsymbol{M}_0 + 2\kappa\boldsymbol{M}_1 + \boldsymbol{M}_2). \tag{32}$$

For second order B-spline basis functions, a similar adjustment can be made to the initial and final 2-by-2 blocks of the matrix. In both cases, Taylor expansion of the boundary correction at a specific $\kappa_0 > 0$ is likely preferable when the temporal construction is applied to the space-time construction in Theorem 4.1.

F. Application details

We performed the computations for the temperature case study on a single node machine with 52 cores (26 dual-socket Intel Xeon Gold 6230R CPU) and 755 GB of main memory. After preliminary model fitting with lower resolution spatial mesh we fitted the model with 1251 mesh nodes. The parallel computations were performed with `inlabru` (version 2.10.0), `INLAspacetime` (version 0.1.7) via `R-INLA` (version 23.11.26) with the `PARDISO` library, using 4 parallel evaluations of the posterior log-density, each one using 8 threads. The average time per function evaluation were 52.94 seconds, 67.96 seconds, 92.06 seconds and 186.05 seconds, respectively for models A, D, C and D. The respective number of evaluations of the posterior density were 479, 1114, 333 and 607, and the total computing time 7.12 hours, 21.09 hours, 8.58 hours and 31.45 hours. The real memory peak were 114.75 GB, 178.78 GB, 219.25 GB and 211.43 GB, respectively. These timings can be significantly reduced by using a different parallelization strategy along with better starting values. Starting values can be set from fits of each one of these models but using lower resolution meshes which allows faster computations.

The computed results were used for the within-sample and leave-one-out prediction scores in Table 3, as well as for the multi-horizon forecast assessment in Section 5.2.4. Details of the multi-horizon forecast scores are shown in Figures 9 and 10, including the mean error (ME, estimated forecast bias), mean absolute error (MAE), mean squared error (MSE), mean Dawid-Sebastiani scores (DS), mean continuous ranked probability score (CRPS), and scale-invariant CRPS (SCRPS).

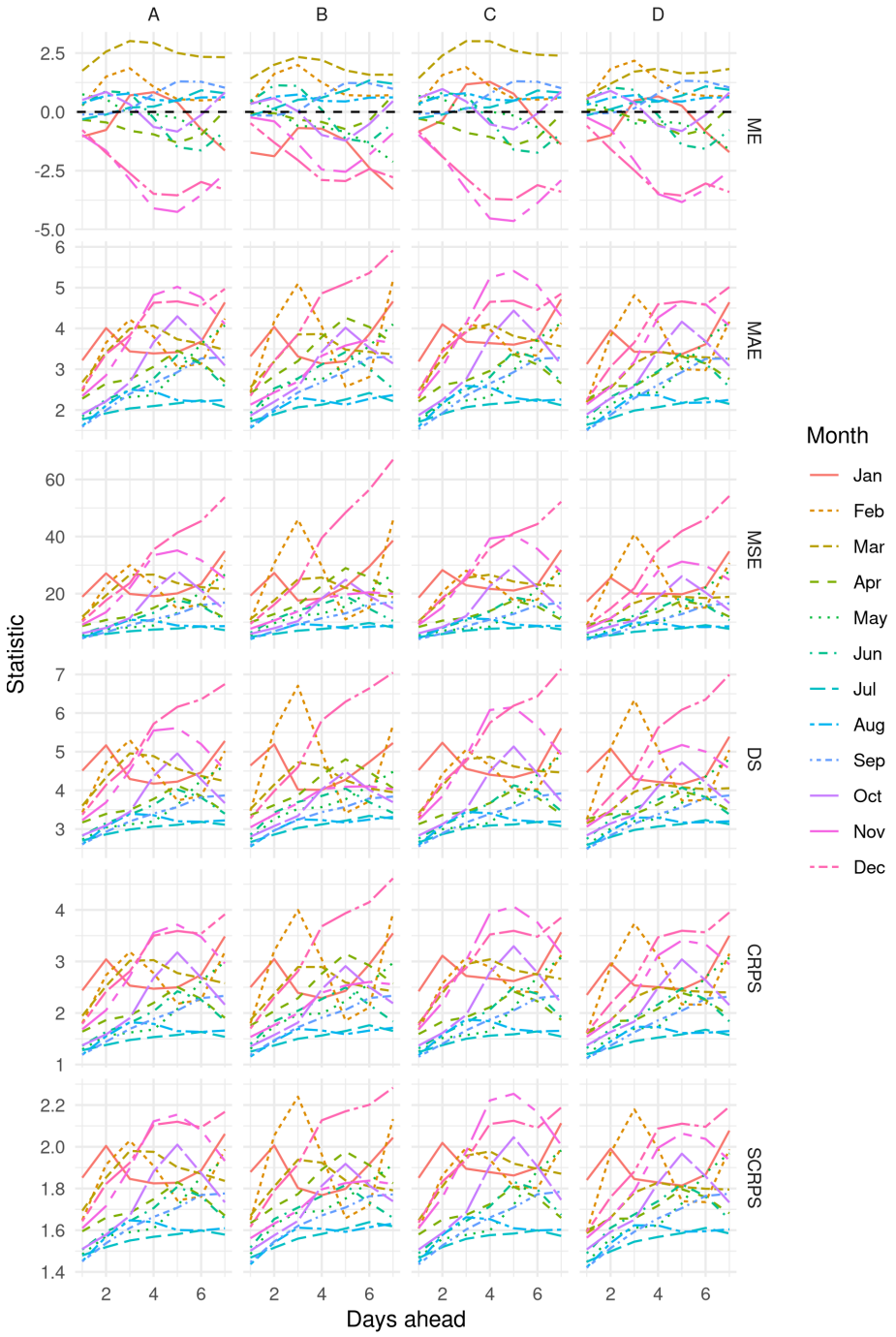


Figure 9. *Mean error and prediction score averages for each model, for each forecast horizon (1–7) and each month of the year, for the multi-horizon multi-scenario setting.*

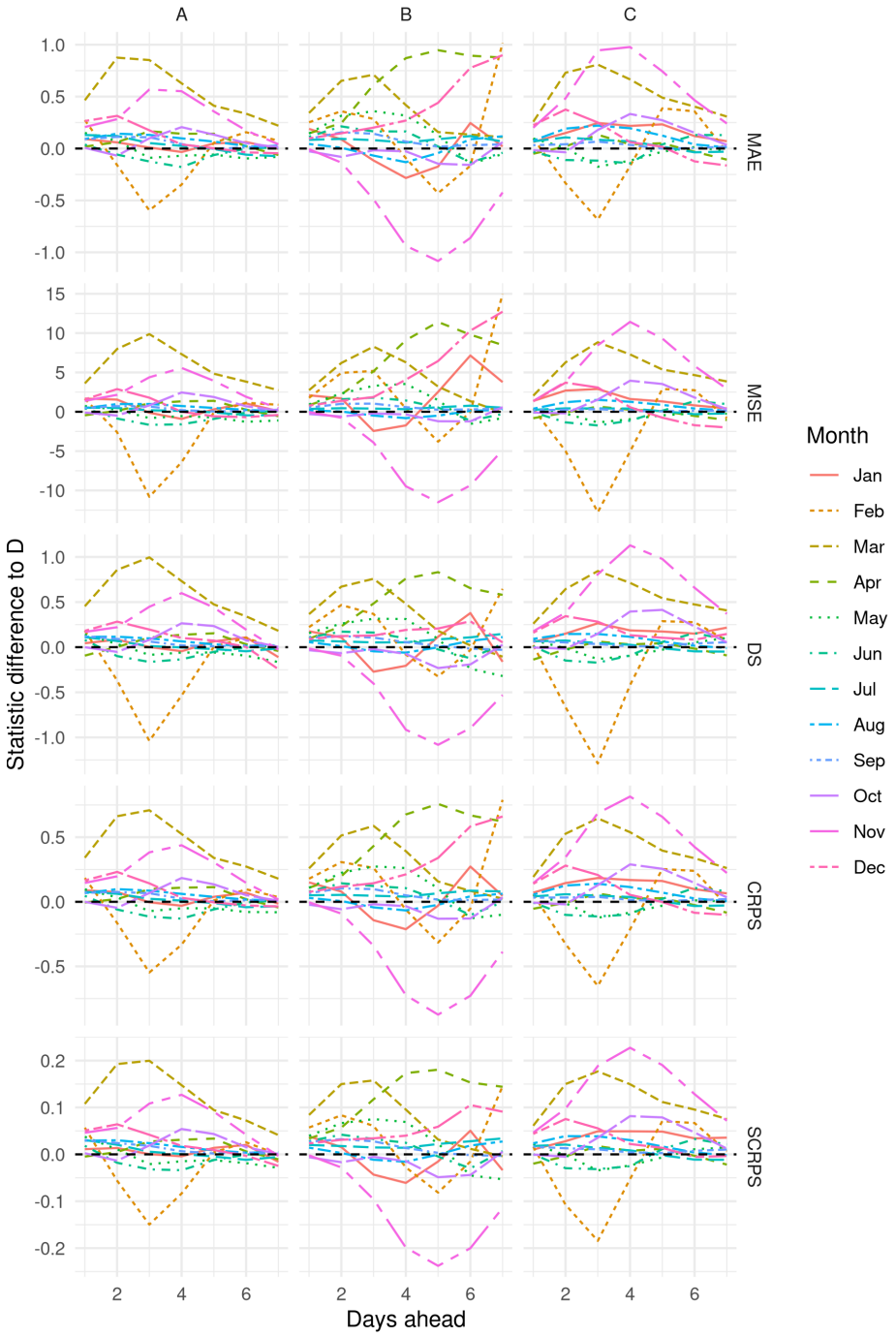


Figure 10. *Prediction score averages for each model with the scores for model D subtracted, for each forecast horizon (1–7) and each month of the year, for the multi-horizon multi-scenario setting.*

Discussion of "A diffusion-based spatio-temporal extension of Gaussian Matérn fields"
by Finn Lindgren, Haakon Bakka, David Bolin, Elias Krainski and Håvard Rue

Christopher K. Wikle
Department of Statistics
University of Missouri, USA

Nathan B. Wikle
Department of Statistics and Actuarial Science
University of Iowa, USA

We congratulate the authors on their spatio-temporal extension of the Matérn spatial process. As outlined in this comprehensive paper, the connection between the Matérn covariance function and its SPDE representation remains of foundational importance to the understanding and construction of geospatial stochastic processes. This paper joins the authors' previous efforts in increasing the awareness (and ease of use) of so-called "constructive" approaches to spatial modeling, most notably their SPDE approach for Gaussian fields (Lindgren, Rue and Lindström, 2011) and its implementation via INLA (Rue et al., 2017). We share the belief that such constructive approaches typically function best when relevant physical processes are incorporated in the model, often to the benefit of the theorist and practitioner alike. Furthermore, the extension from a purely spatial to spatio-temporal model is both elegant and (at least conceptually) simple: specify a dynamical system in the SPDE construction. It is surprising, then, that until Jones and Zhang's (1997) specification of a stationary space-time process via a fractional diffusion equation, this approach received so little attention after its introduction by Heine (1955) and Whittle (1963). Indeed, Jones and Zhang (1997) was ahead of its time; it is nice to see that work getting the attention it deserves and that it has helped to motivate the extension to spatio-temporal Matérn processes presented here. We have little to add in this regard beyond what is said in the paper. Instead, we will focus our brief discussion on the broader SPDE modeling approach.

One interesting question, which might be overlooked upon a first reading of this paper, is whether there is a distinct advantage to specifying the dynamical system in continuous, rather than discrete, time. For example, Cressie and Wikle (2015, Chap. 6) compare the space-time covariance from the diffusion-injection SPDE model of Heine (1955) and the marginal space-time covariance implied by the associated conditional discrete-time dynamic spatio-temporal model (DSTM). Not surprisingly, they are nearly indistinguishable except near the space and time origins. Thus, one obvious advantage of the marginal SPDE approach is the adaptability of the continuous-time model to irregularly observed and/or high-frequency data, although the exact scope of the advantage would depend on the number of observations and the complexity of the specified SPDE. Yet, when the data are reasonably considered to be discrete in time, the conditional DSTM may provide much greater flexibility in deep hierarchical frameworks as it allows parameters

that control the advection, diffusion, and injection processes to themselves be specified as dependent processes (or conditioned on spatio-temporal covariates). This deep conditional modeling, motivated by mechanistic and stochastic parameter processes, is the foundation for the physical-statistical modeling approach pioneered by Mark Berliner and colleagues (e.g., Berliner, 2003; Wikle and Hooten, 2010).

Perhaps a less obvious advantage of continuous SPDE representations in general, and the addition of the temporal smoothness parameter in the model presented here specifically, is its potential adaptability to changes in temporal support. For example, spatial data may represent observations from a time-averaged or aggregated dynamic process (e.g., monthly precipitation totals, annual average pollution concentrations, etc.). In such cases, the SPDE approach can suggest novel marginal spatial dependence structures for the aggregated observations: define the relevant continuous-time process and integrate the solution over the observational time span. This idea has been considered for advection-diffusion processes with certain simplifying assumptions (see Wikle et al. (2022) or Hanks (2017) for a similar approach). However, to our knowledge this has not been attempted with the more general fractional diffusion-like processes considered in this paper. Such an approach may yield relevant marginal spatial dependence structures beyond the fixed-time Gaussian Matérn fields considered here, and continues the broader trend of relating conditional spatial models to relevant spatio-temporal processes found elsewhere in the geospatial literature. We recall D.R. Cox's comment to the classic Besag (1974) work on Markov random fields (MRFs):

> *Nevertheless, understanding of the conditional models may be helped by relating them to temporal-spatial models, and in particular to their stationary distributions. It would be interesting to know what general connections can be established between Mr Besag's auto-models and stationary distributions of simple temporal-spatial processes.* (Besag, 1974, p.225)

Another exciting avenue suggested by the present work concerns extensions to nonlinear spatio-temporal processes. Most biological/environmental/geophysical processes are, at least at some temporal scale, nonlinear. That is, there are explicit interactions between spatio-temporal scales of variability that lead to nonlinear phenomena (density-dependent growth, fronts, predator-prey dynamics, disease spread, etc.). Given that such processes imply higher-order marginal dependence, the constructive approach may motivate new statistical models for such processes. Wikle and Hooten (2010) present a discrete time DSTM approach for quadratic nonlinear processes motivated by PDEs, but the resulting models can be limited by the curse of (parameter) dimensionality and computational tractability. These are the same issues one faces with nonlinear deep neural approaches for modeling nonlinear spatio-temporal processes (see the recent review in Wikle and Zammit-Mangion, 2023). Whether the SPDE approach can overcome these issues is unclear, but it provides, at least, a possible connection (or alternative) to the increasingly popular physically-informed neural network approaches (e.g., Cuomo et al., 2022; Ren et al., 2023).

Another area of future research motivated by the present work is to expand the general manifold ideas herein to suggest new general classes of dependence models for graphical models such as those motivated by time-varying networks (e.g., Ghosh et al., 2022). Indeed, the well-known connection between MRFs and graphical models, and the similar connection between SPDE models and MRF-based implementation methods championed by the authors, suggests a potential to provide mechanistically-motivated process dependence models for time-varying networks, both in their classical representation and in graph neural methods (e.g., Zhou et al., 2020).

In conclusion, we thank the authors for a well-written and important extension of their SPDE/INLA approach to flexible spatio-temporal processes. Not only does this provide a useful suite of models for many real-world data sets, but it has suggested to us several interesting connections and possible extensions to our own work.

References

Berliner, L. M. (2003). Physical-statistical modeling in geophysics, *Journal of Geophysical Research: Atmospheres*, 108.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society: Series B (Methodological)*, 36, 192-225.

Cressie, N. and Wikle, C. K. (2015). *Statistics for Spatio-Temporal Data*, John Wiley & Sons.

Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M. and Piccialli, F. (2022). Scientific machine learning through physics–informed neural networks: Where we are and what's next, *Journal of Scientific Computing*, 92, 88.

Ghosh, D., Frasca, M., Rizzo, A., Majhi, S., Rakshit, S., Alfaro-Bittner, K. and Boccaletti, S. (2022). The synchronized dynamics of time-varying networks, *Physics Reports*, 949, 1-63.

Hanks, E. M. (2017). Modeling spatial covariance using the limiting distribution of spatio-temporal random walks, *Journal of the American Statistical Association*, 112, 497-507.

Heine, V. (1955). Models for two-dimensional stationary stochastic processes, *Biometrika*, 42, 170-178.

Jones, R. H. and Zhang, Y. (1997). Models for continuous stationary space-time processes, *in* P. D. E. R.-C. W. W. R. W. T.G. Gregoire, D.R. Brillinger, ed., *Modelling Longitudinal and Spatially Correlated Data*, Springer, New York, NY, 289-298.

Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73, 423-498.

Ren, P., Rao, C., Liu, Y., Ma, Z., Wang, Q., Wang, J.-X. and Sun, H. (2023). Physr: Physics-informed deep super-resolution for spatiotemporal data, *Journal of Computational Physics*, 492, 112438.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. and Lindgren, F. K. (2017). Bayesian computing with inla: A review, *Annual Review of Statistics and Its Application*, 4, 395-421.

Whittle, P. (1963). Stochastic-processes in several dimensions, *Bulletin of the International Statistical Institute*, 40, 974-994.

Wikle, C. K. and Hooten, M. B. (2010). A general science-based framework for dynamical spatio-temporal models, *Test*, 19, 417-451.

Wikle, C. K. and Zammit-Mangion, A. (2023). Statistical deep learning for spatial and spatiotemporal data, *Annual Review of Statistics and Its Application*, 10, 247-270.

Wikle, N. B., Hanks, E. M., Henneman, L. R. F. and Zigler, C. M. (2022). A mechanistic model of annual sulfate concentrations in the united states, *Journal of the American Statistical Association*, 117, 1082-1093.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C. and Sun, M. (2020). Graph neural networks: A review of methods and applications, *AI open*, 1, 57-81.

Jorge Mateu
Department of Mathematics,
University Jaume I of Castellón, Spain
email: mateu@uji.es

Francesco Serafini
School of Mathematics,
Statistics and Physics, Newcastle
email: francesco.serafini@ncl.ac.uk

The authors are to be congratulated on a valuable and thought-provoking contribution in the rapidly developing field of space-time modelling with tangible grounds in statistics, mathematics and computer sciences. Since the first INLA proposal, thousands of scientific contributions have appeared in the literature, using, adapting and enhancing this Bayesian-based inference framework, starting with spatial contexts, and lately extending this paradigm into space-time in combination with more complex developments. In this context, we underline the undoubtful fact that Gaussian random fields with Matérn covariance functions are not only popular (as written by the authors) but necessary building block models in spatial statistics and machine learning. These two scientific communities are coming together to solve joint problems that time ago were solved separately.

This paper extends the Matérn covariance function to a family of spatio-temporal covariance functions, and the way the authors have chosen their way is through the alternative mathematical representation of stochastic partial differential equations (SPDE), rather than following analytical representations of covariance functions. By doing this, this paper enhances, and builds upon, existing SPDE developments and are able to delineate and define the process with desired properties already encoded in the process.

We acknowledge this diffusion-based extension of the Gaussian Matérn fields to a family of spatio-temporal stochastic processes (DEMF) for a number of reasons, very much in particular for the possibilities of improving existing methodologies in the field of spatio-temporal point processes. Our discussion centres around two main points as follows.

1. Applications to point processes

We have found the sentence *'In practice, however, users of statistical software often choose a model based on convenience. If there are available code examples, the choices made in these will often be carried forward into future analyses."* particularly true in applied contexts, and it highlights the importance of having accessible code for complex

models. An interesting class of models that is not mentioned in the paper but that is already implemented in `inlabru` (Bachl et al., 2019) is the Log-Gaussian Cox Process (LGCP) model. LGCP models are implemented using the methodology described in Simpson et al. (2016) and we believe there is no problem in combining the new class of random effects illustrated in the paper with this point process methodology. Having the ability to use non-separable spatio-temporal GMRFs as random effects in LGCP models is appealing in a number of different applied contexts. In this section, we will provide examples of interesting applications and use of the proposed class of random effects. Most of these examples concern earthquake occurrence but can be used to illustrate different problems.

A first application that comes to mind regarding earthquakes modelling and forecasting is concerned with the magnitude distribution. Two widely popular choices consist in setting a magnitude of completeness M_0, and assuming that the observed magnitudes $m \leq M_0$ follow an exponential distribution; this corresponds to the standard Gutenberg-Richter law (Gutenberg and Richter, 1956), or a truncated Pareto distribution on the seismic moment (Kagan, 1991). In this context, there is great interest in determining whether the parameters of the magnitude distribution are varying over time and space (Herrmann, Piegari and Marzocchi, 2022; El-Isa and Eaton, 2014; Kamer and Hiemer, 2015). These variations are usually estimated partitioning the space-time domain and producing different estimates for each element of the partition. The ability of expressing the parameters as a Gaussian Markov random field (GMRF) with non-separable covariance functions and with a physical interpretation could be very relevant in this context. This will not only be useful in testing more complex hypothesis on the magnitude distribution but also to build better performing models in terms of their operational capabilities (Hiemer and Kamer, 2016). Still in the context of the magnitude distribution, it would be very interesting to express the magnitude threshold M_0 using a spatio-temporal GMRF. This would be important because the magnitude of completeness changes over time and space depending on the quality of the seismographic network and, therefore, it would be appropriate having a non-separable, possibly non-stationary, random effect.

The combination with LGCP models can prove fruitful for modelling long (years, decades) and short (days, weeks, months) term seismicity. Regarding long-term seismicity, Bayliss et al. (2020, 2022) used the SPDE approach and `inlabru` to build models for earthquakes occurrence incorporating covariates and using the random effect to explain the spatial correlation due to unobserved phenomena. Having easy access to non-separable spatio-temporal random effects could greatly improve this type of models. Indeed, these effects would reflect the spatial and temporal change in long-term seismicity, which are mainly due to the variables changing over time and space, such as distance from faults, the level of stress accumulated in the earth crust, and heat flow.

Regarding short-term seismicity models, Serafini, Lindgren and Naylor (2023) adapted the technique used for LGCP to approximate Hawkes process models (Hawkes,

1971), which have intensity of the form

$$\lambda(\mathbf{x}|\mathcal{H}_{\mathbf{x}}) = \mu(\mathbf{x}) + \sum_{\mathbf{x}_i \in \mathcal{H}_{\mathbf{x}}} g(\mathbf{x} - \mathbf{x}_i), \tag{1}$$

where $\mathbf{x}$ is a point in space-time and possibly equipped with one or more marks (e.g., magnitude, depth), and $\mathcal{H}_{\mathbf{x}}$ is the history of the process up to $\mathbf{x}$. The Hawkes process can be seen as the superposition of a background Poisson process $\mu(\mathbf{x})$ representing events occurring spontaneously, and different offspring processes $g(\mathbf{x} - \mathbf{x}_i)$ representing the points induced by observation $\mathbf{x}_i$. In this context, having non-separable, possibly non-stationary, spatio-temporal random effects may be useful in various ways. First of all, similarly to long-term seismicity, it can be used for the background rate $\mu(\mathbf{x})$. This is usually considered constant in time but varying in space and it is factorised as $\mu(\mathbf{x}) = \mu\nu(\mathbf{x})$, where $\nu(\mathbf{x})$ is just the spatial variation integrating to one over the spatial domain. This is usually estimated independently from μ and the parameters of the triggering function ($g(\mathbf{x} - \mathbf{x}_i)$) (Ogata, 2011). Therefore, a natural first extension would be to use an LGCP model for $\nu(\mathbf{x})$ with non-separable spatio-temporal effects to have a spatio-temporally varying background rate. Another important application would be to use such random effects to define the parameters of the triggering function. As before, this is usually done by partitioning the observations. In contrast, having this type of diffusion models would first provide a more mathematically appealing formulation of models with triggering functions that have spatio-temporally varying parameters, and secondly, will likely provide more capable models in terms of forecasting future occurrences. In general, providing accessible code to implement these random effects in combination with LGCP models will enable many applied researchers to formulate and test more complex hypotheses on the earthquake generation process with the potential of shedding light on new aspects of the process, and improve forecasting models used for operational purposes.

Another potential field of application is cosmology and research devoted in studying the evolution of the spatio-temporal correlation between galaxies locations (Kerscher, 2001; Simon, 2007; Friedrich et al., 2021). Indeed, they also use LGCPs to model the locations of galaxies and, therefore, having a non-separable spatio-temporal process would be useful in this context as opposed to the separable models that are in use now. The problems that cosmologists face are similar to the distance sampling problem described on the `inlabru` website[1]. One caveat would be that they usually are interested in covariance functions defined as power laws of the distance between galaxies. For this field, it would be useful to know if there are special settings of the DEMF parameters able to approximate a power law, and if there is the possibility to generalising the approach proposed in the paper to obtain such covariance functions. Even if this would not be possible, providing accessible code for LGCP to be used in combination with DEMF random effects would be beneficial for this field for two reasons, the first one is having the ability to use non-separable models, and secondly, it can leverage the advantages in

[1] https://inlabru-org.github.io/inlabru/articles/2d_lgcp_distancesampling.html

computational time provided by `inlabru`. Indeed, their data are usually composed of petabytes of recordings which usually makes the use of MCMC or similar techniques unfeasible.

2. Non-Euclidean spaces, and anisotropy

Although mathematically very convenient, Euclidean spaces (and Euclidean distances) are arguably not always the best option, and many times they are not even correct in certain contexts. Spatial data living on networks is gaining importance with the development of technology able to provide such type of data. This applies to all types of spatial data, in particular to geostatistical and point pattern data. Linking with the previous section, space-time point pattern models such as LGCPs and Hawkes processes living on networks require accurate while fast inferential frameworks to be able to provide real and helpful solutions to events living on networks.

A number of papers have dealt with the analysis of crime data using self-exciting point process theory, after the analogy drawn by Mohler et al. (2011) between aftershock ETAS models and crime. Zhuang and Mateu (2019) proposed a spatio-temporal Hawkes-type point process model, which includes a background component with daily and weekly periodisation, and a clustering component that is triggered by previous events. However, as crime events are naturally constrained to occur on the streets structure of a city, we advocate the use of point processes on linear networks. A network, or a graph, is a collection of vertices joined by edges. A linear network is a union of finitely many line segments in the plane where different edges only possibly intersect with each other at one of their vertices.

Statistical analysis of network data presents severe challenges (Baddeley et al., 2021). A network is not spatially homogeneous, which creates geometrical and computational complexities and leads to new methodological problems, with a high risk of methodological error. Real network data, as crime data, can also exhibit an extremely wide range of spatial scales. These problems pose a significant challenge to the classical methodology of spatial statistics based on stationary processes, which is largely inapplicable to data on a network. Note also that the choice of distance metric on the network is pivotal in the theoretical development and in the analysis of real data.

As commented above, a Hawkes process can be interpreted as a generalised Poisson cluster process associating to centres, of rate μ, a branching process of descendants. The spatio-temporal Hawkes process has a conditional intensity of the form

$$\lambda(x,y,t) = \mu(x,y,t) + \sum_{t_i;t_i<t} g(x-x_i, y-y_i, t-t_i), \tag{2}$$

where $\mu(x,y,t)$ is the background rate, and $g(x,y,t)$ is the rate of occurrence triggered by an event at time 0 and location at the origin. The triggering density governs the spatio-temporal distance of triggered events from their antecedent events and is usually modelled to decay with distance from the origin over time and space.

Although `inlabru` makes good progress on fitting Hawkes point processes, there is yet a gap in providing a strong framework when network data comes into play. Adapting the diffusion approach proposed in this paper to enhance `inlabru` with fitting capabilities on networks would be welcome by the research community.

Also in this network context, there are some further points to be considered. One is the possibility to generalise equation (24) in the main paper with different basis functions that support processes on non-continuous, non-Euclidean spaces (e.g., processes on networks). Piece-wise linear basis functions, Harmonic basis functions or Karhunen-Loève expansions are still to be extended to these supports, opening theoretical avenues of research. Another aspect is the idea of latent embeddings of point process excitations. When specific events seem to spur others in their wake, marked Hawkes processes enable us to reckon with their statistics. The under-determined empirical nature of these event-triggering mechanisms hinders estimation in the multivariate setting. Spatio-temporal applications alleviate this obstacle by allowing relationships to depend only on relative distances in particular (non-)Euclidean spaces; in this case, we can embed arbitrary event types in a new latent space following the idea of diffusion maps (DM). We might posit a diffusion process across event types. Random walk methods yield approximate manifold embeddings, proven helpful in deep representations. Constructed as graph affinities, the triggering influences guide a Markovian random walk of which diffusion maps may be approximated via spectral decomposition. Indeed, asymmetrical DM embeddings serve as an adequate initial condition, but are not always conducive to stable learning in conjunction with dynamic kernel bases; this approximation builds the DM approach.

A final aspect needed in the context of LGCPs or Hawkes processes is how to exploit inherent properties of existing main directions in the events leading to a clear anisotropic spatial or space-time structure. The paper on discussion focuses on separability and provides a unified framework to deal with separable covariance functions as well as non-separable ones, depending on the value of the parameters. In doing so, the covariance functions they report (e.g. Preposition 3.1, Corollary 3.2.1, Figure 1) are always isotropic. Formulating anisotropic covariance functions is a known exercise and these could drive anisotropic LGCPs. However, using the diffusion-based representation might help in getting deeper the process structure itself while providing comprehensive and flexible anisotropic structures. If, in addition, this can be implemented for network-support point processes, we will be able to handle and model a larger body of more complex problems.

References

Bachl, F. E., Lindgren, F., Borchers, D. L. and Illian, J. B. (2019). inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10, 760-766.

Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G. and Davies, T. (2021). Analysing point patterns on networks – review. *Spatial Statistics*, 42, 100435.

Bayliss, K., Naylor, M., Illian, J. and Main, I. G. (2020). Data-driven optimization of seismicity models using diverse data sets: Generation, evaluation, and ranking using inlabru. *Journal of Geophysical Research: Solid Earth*, 125, e2020JB020226.

Bayliss, K., Naylor, M., Kamranzad, F. and Main, I. (2022). Pseudo-prospective testing of 5-year earthquake forecasts for California using inlabru. *Natural Hazards and Earth System Sciences*, 22, 3231-3246.

El-Isa, Z. H. and Eaton, D. W. (2014). Spatiotemporal variations in the *b*-value of earthquake magnitude-frequency distributions: Classification and causes. *Tectonophysics*, 615, 1-11.

Friedrich, O., Andrade-Oliveira, F., Camacho, H., Alves, O., Rosenfeld, R., Sanchez, J., Fang, X., Eifler, T. F., Krause, E., Chang, C., et al. (2021). Dark energy survey year 3 results: covariance modelling and its impact on parameter estimation and quality of fit. *Monthly Notices of the Royal Astronomical Society*, 508, 3125-3165.

Gutenberg, B. and Richter, C. F. (1956). Earthquake magnitude, intensity, energy, and acceleration: (second paper). *Bulletin of the seismological society of America*, 46, 105-145.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58, 83-90.

Herrmann, M., Piegari, E., and Marzocchi, W. (2022). Revealing the spatiotemporal complexity of the magnitude distribution and b-value during an earthquake sequence. *Nature Communications*, 13, 5087.

Hiemer, S. and Kamer, Y. (2016). Improved seismicity forecast with spatially varying magnitude distribution. *Seismological Research Letters*, 87, 327-336.

Kagan, Y. (1991). Seismic moment distribution. *Geophysical Journal International*, 106, 123-134.

Kamer, Y. and Hiemer, S. (2015). Data-driven spatial b value estimation with applications to california seismicity: To b or not to b. *Journal of Geophysical Research: Solid Earth*, 120, 5191-5214.

Kerscher, M. (2001). Constructing, characterizing, and simulating gaussian and higher-order point distributions. *Physical Review E*, 64, 056109.

Mohler, G. O., Short, M., Brantingham, P., Schoenberg, F. and Tita, G. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106, 100-108.

Ogata, Y. (2011). Significant improvements of the space-time etas model for forecasting of accurate baseline seismicity. *Earth, planets and space*, 63, 217-229.

Serafini, F., Lindgren, F. and Naylor, M. (2023). Approximation of bayesian hawkes process with inlabru. *Environmetrics*, 34, e2798.

Simon, P. (2007). How accurate is limber's equation? *Astronomy & Astrophysics*, 473, 711-714.

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H. and Rue, H. (2016). Going off grid: computationally efficient inference for log-gaussian cox processes. *Biometrika*, 103, 49-70.

Zhuang, J. and Mateu, J. (2019). A semiparametric spatiotemporal hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182, 919-942.

Rejoinder

We thank the discussants for their insightful comments and suggestions. We will address some of the points raised, by topic.

As noted by Wikle & Wikle in their comment, an important aspect of the DEMF models presented in the paper is their physical interpretability. In addition to providing a straightforward method for defining valid space-time stochastic processes without requiring direct access to the resulting covariance function, the continuous definition allows them to form building blocks of complex models. Hierarchical models with multiple observation sources and types of data are easily defined, and the same hierarchical model can involve point-referenced measurements as well as spatially and temporally aggregated information. Moreover, just as in the spatial case handled in Simpson et al. (2016) and extended by Bachl et al. (2019), the processes can be used as building blocks for log-Gaussian Cox process observation models, by constructing numerical integration schemes based on the basis functions used to numerically represent or approximate the theoretical process realisations.

In their comment, Wikle & Wikle also raise the question of whether the increased flexibility of time-discrete models may be advantageous. While such flexibility can be useful in specific situations where the processes exhibit variation and structure on a shorter time scale than the numerical representation can handle in a continuous-time approach, in general we believe that the greater interpretability of time-continuous models outweighs this; see also Simpson, Lindgren and Rue (2011) for a related discussion. One reason is that some of the flexibility of valid time-discrete models constitutes non-physical behaviour that is sensitive to changes in the temporal resolution. In addition, the possibilities for time-discrete models that do have physical meaning, such as advection, diffusion, and injection processes, are equally applicable in the time-continuous setting. The only real obstacle to such extensions, including allowing the model parameter themselves to be spatial, temporal, or spatio-temporal processes, is the practical numerical implementation and inference, both in terms of computational cost and the need for sufficient amounts of information in the available data. Indeed, the comment mentions that the curse of dimensionality also affects the time-discrete approaches. In practice, the modeller needs to decide what temporal spatial resolution, and structural flexibility is needed for any given problem.

Both discussion comments bring up the idea of extending the models beyond $\mathbb{R}^d$ and $\mathbb{S}^2$. One advantage of the SPDE approach to defining stochastic processes and using finite elements or related methods for practical computation is that extensions to curved spaces is relatively straightforward. For sufficiently smooth manifolds, much of the theory and practical implementation details remain the same. The only real difference is the need to define the differential operators in terms of the metric of the manifold. For purely

spatial fields, this was covered for Whittle-Matérn fields in Lindgren, Rue and Lindström (2011), allowing models to be defined on curved compact manifolds (Coveney et al., 2020). As can be seen by the proof of the precision structure construction in the paper, both non-stationary and manifold extensions of the DEMF models are straightforward, as each term in the precision matrix sum only involves Kronecker products of separable time and space operators; See Kirchner and Willems (2023) for recent theoretical results. We agree that it would be interesting to consider extensions of the proposed models to networks or graphs, and this is a topic that we are currently investigating, and we discuss this in further detail below.

In their comments, both Mateu & Serafini and Wikle & Wikle touch upon different aspects of either extending the models to non-linear behaviour or to use them as building blocks in other models, such as Hawkes point process. Continuing the theme of taking inspiration from physics, we note that even in the deterministic case, non-linear PDE models can be extremely challenging, both from a theoretical and practical point of view. For example, the Navier-Stokes equations still lack a general existence and smoothness theory. However, a potentially fruitful avenue may be to extend linearisation techniques to the stochastic setting, by considering linear SPDEs whose solutions approximate the statistical properties of the original non-linear models.

When it comes to using the models as building blocks, such as modelling spatio-temporally varying parameters of a Hawkes process excitation kernel, the need for non-separable models is less clear. In point process settings, the observations generally are only weakly informative about the model parameters, and the non-separability would likely only be a higher-order effect that cannot be reliably estimated or used, and that covariance product separable models are likely to be sufficient. The main utility of the DEMF models is likely to be as primary models in a hierarchical model where the parameters are modelled in a more parsimonious way, ideally in combination with observed covariates with at least qualitatively known impact on the processes.

We agree with Mateu & Serafini that there are exciting opportunities to create new models by finding SPDE operators that generate fields with other properties than those of the Whittle-Matérn class, such as power law covariance functions. Spatially oscillating fields have already been constructed without direct reference to the wave equation Lindgren et al. (2011), but directly adapting the wave equation could have useful applications and increased interpretability, including intermediate models between the heat and wave equations. Another possibility is to exploit the same technique as used by Bolin, Simas and Xiong (2023) and Sørbye, Myrvoll-Nilsen and Rue (2019) to approximate Whittle-Matérn models with fractional exponents. In this approach, the target model is approximated by a sum of just a few high-order Markov processes, but the resulting processes are non-Markovian.

We now turn our attention to the issue of how to define space-time processes on networks or graphs. A suitable space to introduce this extension on is a metric graph Γ, which is defined in terms of a set of vertices $\mathcal{V}$ and a set of edges $\mathcal{E}$ connecting the vertices. The difference to a regular graph is that the edges are defined as rectifiable

curves, and a position $s \in \Gamma$ can be represented as (e,t), where $e \in \mathcal{E}$ denotes an edge, and t is a position on that edge. Thus, these spaces contain linear networks as a special case.

Recently, we introduced Whittle–Matérn fields on metric graphs as the solution to

$$(\kappa^2 - \Delta_\Gamma)^{\alpha/2}(\tau u) = \mathcal{W}, \quad \text{on } \Gamma \tag{1}$$

where $\alpha > 1/2$, $\mathcal{W}$ is Gaussian white noise on Γ and Δ_Γ is the so-called Kirchhoff Laplacian on Γ, which is an operator that acts as the second derivative on the edges (Bolin, Simas and Wallin, 2023a).

This model is well-posed for any compact metric graph, and α controls the sample path regularity of the solutions in the same way as it does when the model is posed on Euclidean domains. Thus, the model can be used to define differentiable Gaussian processes on general metric graphs, and this is as far as we know, the only construction that can do so. Further, if $\alpha \in \mathbb{N}$, these models are Markov random fields (Bolin, Simas and Wallin, 2023b) and this can be used to perform exact and computationally efficient likelihood-based inference (Bolin, Simas and Wallin, 2023d). For fractional α, and for generalised Whittle–Matérn fields where κ and τ are spatially varying, FEM approximations can be used to obtain computationally efficient approximations (Bolin et al., 2023).

These spatial models are implemented in the R package MetricGraph (Bolin, Simas and Wallin, 2023c), which also contains an implementation of LGCPs on metric graphs. An example of a simulated point pattern based on a Whittle–Matérn field with $\alpha = 2$, and an estimate of the log-intensity using R-INLA can be seen in Figure 1.

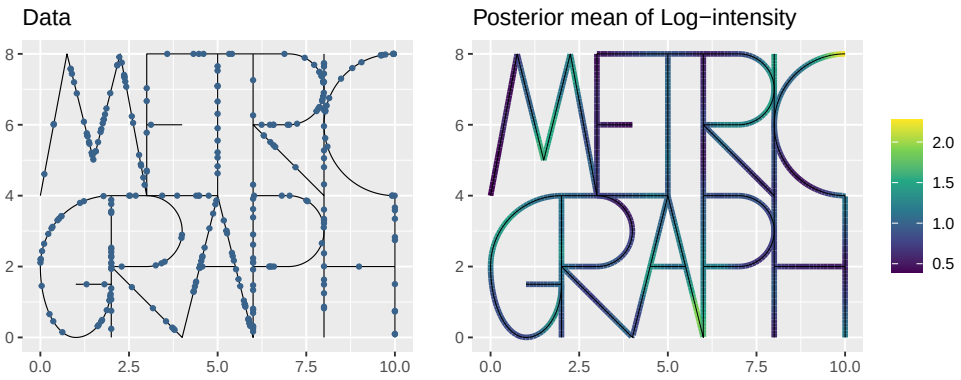


Figure 1. *Simulated LGCP on a metric graph and estimate of log-intensity based on R-INLA and the MetricGraph package.*

We are working on the extension of the spatio-temporal models introduced in this work to the setting of metric graphs. In this case, it is important to consider an extension

of the model which includes an advection term,

$$(\gamma_t \frac{d}{dt} + (\kappa^2 + \rho\, \mathrm{d}_s - \Delta_\Gamma)^{\alpha_s/2})^{\alpha_t} u = \mathrm{d}\mathcal{E}_Q, \quad \text{on } \Gamma \times [0,T], \tag{2}$$

because many datasets on metric graphs, such as river systems, have a clear transport direction. It should be noted that even though the spatio-temporal models considered in this work result in isotropic covariances when posed on $\mathbb{R}^d$, this is not the case for the metric graph setting: Even if $\rho = 0$ in (2), the model is anisotropic on general metric graphs. Thus, we agree that isotropy typically is not a realistic property on metric graphs.

Finally, there are indeed close connections between graphical models and Whittle–Matérn fields on metric graphs, because certain models based on the graph Laplacian can be viewed as finite difference approximations to (1) (Bolin et al., 2023d). Finding similar connections between the spatio-temporal model (2) and spatio-temporal graphical models is an interesting topic for future work.

References

Bachl, F. E., Lindgren, F., Borchers, D. L. and Illian, J. B. (2019). inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10, 760-766.

Bolin, D., Kovács, M., Kumar, V. and Simas, A. B. (2023). Regularity and numerical approximation of fractional elliptic differential equations on compact metric graphs. *Math. Comp.* In press.

Bolin, D., Simas, A. B. and Wallin, J. (2023a). Gaussian Whittle-Matérn fields on metric graphs. *Bernoulli*. In press.

Bolin, D., Simas, A. B. and Wallin, J. (2023b). Markov properties of Gaussian random fields on compact metric graphs. Preprint, arxiv:2304.03190.

Bolin, D., Simas, A. B. and Wallin, J. (2023c). *MetricGraph: Random fields on metric graphs*. R package version 1.1.2, https://CRAN.R-project.org/package=MetricGraph.

Bolin, D., Simas, A. B. and Wallin, J. (2023d). Statistical inference for Gaussian Whittle–Matérn fields on metric graphs. Preprint, arXiv:2304.10372.

Bolin, D., Simas, A. B. and Xiong, Z. (2023). Covariance-based rational approximations of fractional SPDEs for computationally efficient Bayesian inference. *J. Comput. Graph. Statist.* In press.

Coveney, S., Corrado, C., Roney, C. H., Wilkinson, R. D., Oakley, J. E., Lindgren, F., Williams, S. E., O'Neill, M. D., Niederer, S. A. and Clayton, R. H. (2020). Probabilistic interpolation of uncertain local activation times on human atrial manifolds. *IEEE transactions on bio-medical engineering*, 67, 99-109.

Kirchner, K. and Willems, J. (2023). Regularity theory for a new class of fractional parabolic stochastic evolution equations. *Stochastics and Partial Differential Equations: Analysis and Computations*.

Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation ap-

proach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 423-498.

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H. and Rue, H. (2016). Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103, 49-70.

Simpson, D. P., Lindgren, F. K. and Rue, H. (2011). Think continous: Markovian Gaussian models in spatial statistics. *Spatial Statistics*, 1, 16-29.

Sørbye, S. H., Myrvoll-Nilsen, E. and Rue, H. (2019). An approximate fractional Gaussian noise model with $O(n)$ computational cost. *Statistics and Computing*, 29, 821-833.

SORT 48 (1) January-June 2024, 67-92 DOI: 10.57645/20.8080.02.14

Estimation of logistic regression parameters for complex survey data: simulation study based on real survey data

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

Abstract

In complex survey data, each sampled observation has assigned a sampling weight, indicating the number of units that it represents in the population. Whether sampling weights should or not be considered in the estimation process of model parameters is a question that still continues to generate much discussion among researchers in different fields. We aim to contribute to this debate by means of a real data based simulation study in the framework of logistic regression models. In order to study their performance, three methods have been considered for estimating the coefficients of the logistic regression model: a) the unweighted model, b) the weighted model, and c) the unweighted mixed model. The results suggest the use of the weighted logistic regression model is superior, showing the importance of using sampling weights in the estimation of the model parameters.

MSC: *62J12, 62P25, 62D05.*

Keywords: *Complex survey data, sampling weights, logistic regression, estimation of model parameters, real data based simulation study.*

1. Introduction

Complex survey data are increasingly used by researchers and analysts from different fields. In particular, complex survey data are usually used, among other purposes, to fit

[*] *Corresponding author:* amaia.iparragirre@ehu.eus. Address: Departamento de Matemáticas. Facultad de Ciencia y Tecnología. Universidad del País Vasco (UPV/EHU). Barrio Sarriena s/n. 48940 Leioa.

[1] Departamento de Matemáticas, Universidad del País Vasco (UPV/EHU).

[2] Eustat - Euskal Estatistika Erakundea - Instituto Vasco de Estadística.

[3] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain.

Received: August 2022
Accepted: May 2023

prediction models. These data are commonly obtained by sampling the finite population that is of interest for the survey by some complex sampling design. One of the characteristics of this type of data are sampling weights, which indicate the number of units that each sampled observation represents in the finite population. When working with complex survey data, before implementing the traditional statistical techniques, most of which have been designed to be implemented on simple random samples, it should be assessed whether these techniques are valid to this kind of data (Skinner, Holt and Smith, 1989).

In particular, whether or not to use the sampling weights when fitting prediction models is a question that has been widely discussed in the literature by a number of researchers (Brewer and Mellor, 1973; Smith, 1981). Different perspectives can be adopted when fitting prediction models to survey data, which are usually denoted as model -and design-based approaches (Binder and Roberts, 2009; Chambers and Skinner, 2003). On the one hand, the researchers that adopt the design-based perspective warn that if the complex sampling design, and in particular, the sampling weights are not considered in the estimation process of model parameters, the variances tend to be underestimated and biased estimates may be obtained (Binder and Roberts, 2009; Heeringa, West and Berglund, 2017). Therefore, they claim that the sampling weights should be considered in the estimation process of model parameters.

On the other hand, from a model-based point of view, if the model is well specified the coefficient estimates must be unbiased even though the sampling weights are not considered directly in the estimation process and considering them may increase the standard deviations of the estimates, particularly for small sample sizes (Scott and Wild, 1986; Reiter et al., 2005; Chambers and Skinner, 2003; Korn and Graubard, 1995). In this context, Rubin (1976); Scott (1977); Sugden and Smith (1984) established conditions under which the sampling design may be ignored for inference purposes. As explained by Skinner et al. (1989) a condition for a design to be ignorable is to be noninformative. A sampling design is denoted as informative if the response variable is related to the sampling weights, even after considering the covariates that are going to be part of the model (Pfeffermann and Sverchkov, 2009). Different methods have been proposed from the model-based perspective in order to ensure that the design is ignorable and the models are well specified (Pfeffermann and Sverchkov, 2009). Researchers that adopt this perspective propose, among other techniques, to incorporate into the model as covariates all the design variables that have been considered in the sampling process and the interactions between them (see, e.g., DeMets and Halperin (1977); Nathan and Holt (1980); Gelman (2007)).

Although it was already pointed out by Chambers and Skinner (2003), the discussion between the two perspectives is still alive. Some more recent works, such as Reiter, Zanutto and Hunter (2005), Masood, Newton and Reidpath (2016) and Lumley and Scott (2017), show that this debate still generates doubts among researchers and makes it difficult to decide whether or not to use sampling weights in their analyses. Most researchers agree that it is not advisable to ignore sampling weights if the sample is informative

or the model is not well specified, but at the same time, they encourage to ignore the sampling weights when they are not strictly necessary. The difficulty usually lies in identifying whether or not sampling weights are necessary to estimate model parameters based on our particular survey data, or put it another way, whether or not the design is informative. As explained by Pfeffermann and Sverchkov (2009), informativeness depends not only on the sampling design, but also on the model that is going to be fitted, the response variable of that model and the covariates that will be included. Therefore, commonly it is not easy to know whether the sampling design of the survey data to be analysed is informative or not to fit a particular model. In addition, it is not always possible to include all the design variables and the interactions between them in the model due to several reasons, such as the lack of information, the large number of design variables and the fact that when including design variables as covariates into the model it may lose scientific interpretability (Pfeffermann and Sverchkov, 2009). Consequently, nowadays, it is not easy to decide in practice whether sampling weights should or not be considered for estimating model parameters yet. For this reason, we believe that further studies are needed in this area and, in particular, we consider that it is necessary to provide insight considering real data based simulation studies, as a complement to the theoretical results and case studies that have been most discussed so far.

Throughout this work we focus on the estimation of model parameters and, in particular, on the logistic regression framework for dichotomous response variables. Although in general there are more studies concerned with the linear regression model (see, e.g., DeMets and Halperin (1977); Nathan and Holt (1980); Holt, Smith and Winter (1980); Hausman and Wise (1981)), a number of works have also been carried out in order to address this problem arising from complex samplings in the field of logistic regression models. In particular, Scott and Wild (1986, 2002) work with simulated data inspired from a case-control study. It should be noted that case-control studies consist in stratifying the data based on the dichotomous response variable, and therefore, are always based on informative sampling designs. But, what if we do not know whether our sampling design is informative or not to fit a certain model? As mentioned above, in practice, this is the situation that usually occurs when working with real complex survey data. Chambless and Boyle (1985); Lumley and Scott (2017) and Reiter et al. (2005) raise this issue in their analysis with real survey data and they compare several estimation methods adopting both, model- and design-based perspectives and they finally select the most appropriate model for their analysis. However, how can we know in practice whether these differences in estimates are large or not, and if so, which of the estimates is the most appropriate? In this work we aim to go a step further and contribute to the work that has been done in the above-mentioned papers by analysing the differences among different methods by means of a real data based simulation study, in order to work under a real-life scenario that allows us to compare the coefficient estimates to the theoretical ones. Hence, data were generated based on real surveys and, a priori, whether these data are informative or not to fit different models it is not known for us in advance. Our goal is to analyse by means of a simulation study a situation that frequently occurs in

practice and to analyse and evaluate the consequences or the effect of making the decision to consider or not the sampling weights to estimate the coefficients of the logistic regression model in each situation. In this study we compare the performance of several estimation methods that are commonly applied for estimating the coefficients of the logistic regression model (see, for example, Lumley and Scott (2017)). In particular, we compare the coefficient estimates obtained by: a) the unweighted logistic regression model, b) the weighted logistic regression model, and c) the unweighted logistic regression mixed model with random intercept. Different scenarios were defined based on a) data obtained from two different real surveys; and b) number of covariates/parameters in the model. The real surveys were designed and collected by the Official Statistics Basque Office (Eustat) based on single-stage stratification with simple random sampling in each stratum.

The rest of the document is organized as follows. In Section 2 we describe the two original real surveys that motivated this work: ESIE and PRA surveys. In Section 3 the methods that were applied for estimating the model parameters are described. Information about the simulation procedure, scenarios that were drawn and the results we obtained can be found in Section 4. In Section 5, we apply the described methods to real survey data for illustration purposes. Finally, the paper concludes with a discussion in Section 6.

2. Motivating data sets

In this section, we describe the two complex surveys that motivated this work. These surveys were designed and conducted by the Official Statistics Basque Office (Eustat).

On the one hand, the Information Society Survey[1] in companies, which is usually denoted as ESIE survey for its Spanish acronym, was carried out among the companies in the Basque Country in order to collect information about the use of technology. In particular, the response variable that we concern about in this study is a dichotomous response variable that indicates whether a company has its own web-page (1) or not (0), which we aim to model by means of covariates such as ownership (which indicates whether the company is a corporation, limited liability company, public administration,...), activity or number of employees of the establishment. On the other hand, the Population in Relation to Activity (PRA) Survey[2] was conducted among the inhabitants of the Basque Country aged 16 and over, with the aim of estimating the percentage of the labor force of the Basque Country. Specifically, the response variable that we consider in this study indicates whether each individual is active (1) or not (0). Among the most important covariates were age, educational level, nationality, and sex.

In both surveys, the two finite populations were sampled based on single-stage stratification with simple random sampling in each stratum, i.e., the populations were split into different strata, and a certain number of units (that were previously determined)

[1] https://en.eustat.eus/estadisticas/tema_150/opt_1/tipo_7/temas.html

[2] https://en.eustat.eus/estadisticas/tema_37/opt_0/temas.html

were sampled randomly from each stratum. Nevertheless, the strata were defined in very different ways in both surveys. In the ESIE survey, strata were defined based on the combination of three categorical variables which are 1) province where the company is located (that takes 3 categories), 2) activity of the company (in 65 categories) and 3) number of employees (3 categories). Therefore, a large amount of small strata, a total of 585 were defined. However, it should be noted that in some of these strata there are no units in the population, so in fact we have 515 strata in total. In contrast, in the PRA survey, strata are the 23 regions of the Basque Country. This causes the response variable to be distributed differently in each stratum in ESIE, while in PRA, there are no differences of the distribution of the response variable among the strata. In both, ESIE and PRA surveys, once the sample was obtained from the finite population following the described sampling process, a sampling weight was assigned to each sampled unit.

In the ESIE survey, from the finite population of 195 222 companies, 7 725 were sampled (these data was collected in 2010). In particular, strata sizes in the finite population range from 1 to 14 535, where the median is 38 and the interquartile range $7-185.5$. The sampling probabilities for each stratum range from 0.0061 to 1, with a median of 0.2830 and an interquartile range of $0.0970-0.8417$. In contrast, in the PRA survey, from a total of 1 851 316 individuals 10 609 were sampled (information related to the last quarter of 2016). Specifically, strata sizes range from 2 768 to 438 595, being the median 44 335 and $22\,024-72\,834$ the interquartile range. The sampling probabilities range from 0.0041 to 0.0488, with a median of 0.0063 (the interquartile range is $0.0055-0.0102$).

3. Methods

In this section, we describe the methods we have considered in order to estimate the logistic regression coefficients for complex survey data.

Let Y indicate the dichotomous response variable, which takes the value 1 to indicate the event of interest (0 otherwise), and $\boldsymbol{X} = (X_1,\ldots,X_p)^\intercal$ the vector of p covariates. Let $U = \{1,\ldots,N\}$ be a finite population for which N realizations of the set of random variables $(Y,\boldsymbol{X})$ are associated, i.e., $\{(y_i,\boldsymbol{x}_i)\}_{i=1}^N$. Let S be a sample of n observations drawn from the finite population U by single-stage stratification. Let $h = 1,\ldots,H$ indicate the different strata. The sampling weights associated to each sampled unit $i \in S$ are denoted as w_i.

Let us define the true population logistic regression model as follows:

$$logit(p_i) = \ln\left[p_i/(1-p_i)\right] = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}^{True} \tag{1}$$

where $p_i = P(Y=1|\boldsymbol{x}_i)$ denotes the probability of event for the unit i given the values of covariates $\boldsymbol{x}_i$ $(\forall i \in U)$ and the model coefficients $\boldsymbol{\beta}^{True} = (\beta_0^{True},\beta_1^{True},\ldots,\beta_p^{True})^\intercal$ are computed by maximizing the population likelihood:

$$L_{pop}(\boldsymbol{\beta}) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}. \tag{2}$$

However, it should be noted that responses y_i are usually known only for the sampled units, $i \in S$. For this reason, the model should be estimated based on the sample S. In this work, we compare the performance of several estimation methods that are commonly applied for estimating the coefficients of the logistic regression model for dichotomous response variable (Lumley and Scott, 2017). The goal is to compare these estimates to $\boldsymbol{\beta}^{True}$ in order to analyse the performance of each method.

In this context, a simple logistic regression model can be fitted to the complex survey sample S, which can be defined as follows:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}. \tag{3}$$

Different methods can be applied to estimate the vector of regression coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\mathsf{T}}$ based on S:

M1. Unweighted logistic regression model

This method consists in estimating the model coefficients by maximizing the likelihood function in equation (4) by means of some iterative algorithms such as the iteratively reweighted least squares (IRLS) algorithm (McCullagh and Nelder, 1989):

$$L(\boldsymbol{\beta}) = \prod_{i \in S} p_i^{y_i} (1-p_i)^{1-y_i}. \tag{4}$$

Let us denote as $\hat{\boldsymbol{\beta}}_{\text{M1}}$ the coefficients estimated by means of the maximum likelihood method, hereinafter.

M2. Weighted logistic regression model

This approach consists in estimating the coefficients that maximizes the pseudo-likelihood function in equation (5) (Binder, 1981, 1983) which considers the sampling weights w_i:

$$PL(\boldsymbol{\beta}) = \prod_{i \in S} p_i^{y_i w_i} (1-p_i)^{(1-y_i)w_i}. \tag{5}$$

The pseudo-likelihood function is also maximized by means of iterative algorithms (Heeringa et al., 2017; Wolter, 2007). Let us denote as $\hat{\boldsymbol{\beta}}_{\text{M2}}$ the coefficient estimates obtained based on this method.

In addition to the above-mentioned methods, another option is to fit a mixed model considering the complex sampling design as second level units (see, e.g. Lumley and Scott (2017); Masood et al. (2016)). In this study, in particular, we consider a random intercept model in the same way as in Lumley and Scott (2017). Let $i = 1, \ldots, n_h$ indicate the sampled units belonging to stratum h ($\forall h \in \{1, \ldots, H\}$), while $\boldsymbol{x}_{hi}$ and y_{hi} indicate the values of the vector of covariates and response variable for i in stratum h, respectively. Then, we aim to fit the following model to our sample S:

$$\text{logit}(p_{hi}) = \ln\left(\frac{p_{hi}}{1-p_{hi}}\right) = \boldsymbol{x}_{hi}^{\mathsf{T}}\boldsymbol{\gamma} + u_h, \quad u_h \sim N(0, \sigma_u^2). \tag{6}$$

where $p_{hi} = P(Y = 1|\boldsymbol{x}_{hi}, u_h) = \dfrac{e^{\boldsymbol{x}_{hi}^{\mathrm{T}}\boldsymbol{\gamma}+u_h}}{1+e^{\boldsymbol{x}_{hi}^{\mathrm{T}}\boldsymbol{\gamma}+u_h}}$.

M3. Unweighted logistic regression model with random intercept

In this case, the likelihood function is defined as follows:

$$L_{\mathrm{mix}}(\boldsymbol{\gamma}, \sigma_u^2) = \prod_{h=1}^{H} \int_{-\infty}^{+\infty} f(y_{hi}|\boldsymbol{x}_{hi}, u_h) f(u_h) du_h, \tag{7}$$

where $f(y_{hi}|\boldsymbol{x}_{hi}, u_h) = \prod_{i=1}^{n_h} p_{hi}^{y_{hi}}(1-p_{hi})^{1-y_{hi}}$ and $f(u_h) = \dfrac{1}{\sigma_u\sqrt{2\pi}} e^{-u_h^2/2\sigma_u^2}$. The parameters $\boldsymbol{\gamma}$ and σ_u^2 are commonly estimated by maximizing the likelihood function in (7) numerically, usually by means of Laplace approximation (Lee and Nelder, 2001). Let us denote as $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}_u^2$ those estimates, respectively, hereinafter.

However, the comparison of the coefficients obtained from conditional random effect models and the corresponding marginal models is not straightforward (Lee and Nelder, 2004). In the case of logistic random intercept models, marginal coefficients $\boldsymbol{\beta}$ can be obtained based on conditional parameters $\boldsymbol{\gamma}$ as follows:

$$\boldsymbol{\beta} = \frac{\boldsymbol{\gamma}}{\sqrt{1+c^2\sigma_u^2}}, \tag{8}$$

where $c = (16\sqrt{3})/(15\pi)$ (Diggle, Liang and Zeger, 2002). Let us denote as $\hat{\boldsymbol{\beta}}_{\mathrm{M3}}$ the coefficient estimates obtained based on $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}_u^2$.

The goal is to analyse the performance of the above-mentioned methods by comparing the estimates $\hat{\boldsymbol{\beta}}_{\mathrm{M1}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{M2}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{M3}}$ to the true finite population coefficients $\boldsymbol{\beta}^{True}$.

4. Simulation study

In this section, we describe the simulation study that we have conducted in order to analyse the behaviour of the estimation methods described in Section 3 for estimating the coefficients of the logistic regression model based on complex survey data under different scenarios. As mentioned previously, our goal in this study is to compare the coefficient estimates to the true finite population coefficients in real data-based scenarios.

In Section 4.1 the simulation process is described in detail and in Section 4.2 the results obtained in the simulation study are shown.

4.1. Scenarios and set up

In this section, we describe the different scenarios where the simulation study has been conducted and the steps we have followed. The simulation process is described below, step by step:

Step 1. Generate the pseudo-population U of N units from the set of random variables $(Y,\boldsymbol{X})$ (see Appendix A): $\{(y_i,\boldsymbol{x}_i)\}_{i=1}^{N}$.

Step 2. Compute $\boldsymbol{\beta}^{True}$ by maximizing the population likelihood in (2).

For $r = 1,\ldots,R$ repeat the following steps:

Step 3. Obtain a sample $S^r \subset U$ by single-stage stratified sampling and assign the corresponding sampling weights $w_i, \forall i \in S^r$ (see Appendix B).

Step 4. Fit the models to S^r by the likelihood functions in (4), (5) and (7) and obtain $\hat{\boldsymbol{\beta}}^{r}_{\mathrm{M1}}$, $\hat{\boldsymbol{\beta}}^{r}_{\mathrm{M2}}$ and $\hat{\boldsymbol{\beta}}^{r}_{\mathrm{M3}}$, respectively.

Finally, for the results obtained based on samples $r = 1,\ldots,R$ and for each method $\forall m \in \{\mathrm{M1},\mathrm{M2},\mathrm{M3}\}$, let us define the bias of the coefficient vector estimates as follows:

$$\mathrm{bias}_d^r = \hat{\beta}_{d,m}^r - \beta_d^{True}, \quad \forall d = 0,1,\ldots,p. \tag{9}$$

Then, the average bias (AvBias) and the mean squared error (MSE) across $\forall r = 1,\ldots,R$ are defined in equations (10) and (11), respectively:

$$\mathrm{AvBias}_d = \frac{1}{R}\sum_{r=1}^{R}(\mathrm{bias}_d^r) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_{d,m}^r - \beta_d^{True}\right), \quad \forall d = 0,1,\ldots,p, \tag{10}$$

$$\mathrm{MSE}_d = \frac{1}{R}\sum_{r=1}^{R}(\mathrm{bias}_d^r)^2 = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_{d,m}^r - \beta_d^{True}\right)^2, \quad \forall d = 0,1,\ldots,p. \tag{11}$$

Two scenarios have been defined based on the two real surveys described in Section 2, ESIE (Scenario 1, hereinafter) and PRA (Scenario 2, hereinafter). One finite pseudo-population was generated based on each of the surveys (described in **Step 1.**, see Appendix A). Those populations were sampled based on the complex sampling designs that were applied by Eustat in the corresponding real surveys (defined in **Step 3.**, see Appendix B). A total of $R = 500$ samples were obtained from each pseudo-population.

In addition, two different models were fitted to the finite population as well as to the samples for each of the surveys with different number of covariates (**Step 2**). In particular, in Scenario 1 models with $p = 1$ (X_1) and $p = 3$ (X_1, X_2 and X_3) covariates were fitted. In the same way, in Scenario 2, the models were fitted with $p = 1$ (X_1) and $p = 4$ (X_1, X_2, X_3 and X_4) covariates.

It should be noted that all the covariates are categorical and one coefficient was estimated for each category, except for the one considered as reference category. In particular, in Scenario 1, a total of $l = 7$ parameters (including the intercept, β_0) are estimated for the model with $p = 1$ covariates and $l = 14$ parameters for $p = 3$. In the same way, in Scenario 2, $l = 7$ parameters are estimated for $p = 1$ and $l = 14$ parameters for $p = 4$.

All computations were performed in (64 bit) R 4.0.5 (R Core Team, 2021) and a workstation equipped with 32GB of RAM, an Intel i7-8700 processor (3.20 Ghz) and

Windows 10 operating system. In particular, the unweighted logistic regression models (M1) were fitted by means of the `glm` function from the `stats` package, the weighted logistic regression models (M2) by means of the `svyglm` from the `survey` package (Lumley, 2019) and the unweighted mixed models with random intercept (M3) by the `glmer` of the `lme4` package (Bates et al., 2015).

4.2. Results

In this section, we describe the results we obtained in both scenarios: Scenario 1 (which is based on the ESIE survey) and Scenario 2 (which is based on the PRA survey). As explained in Section 4.1, in each scenario two models were fitted with different number of covariates. Our goal is to compare the estimates obtained based on the three coefficient estimation methods described in Section 3 (which are the unweighted logistic regression (M1), the weighted logistic regression (M2) and the unweighted logistic regression with random intercept (M3)) to the true finite population coefficients ($\boldsymbol{\beta}^{True}$), in terms of bias and MSE.

Due to the large number of results obtained, we begin by summarizing the main findings. When comparing the performance of the three methods in each scenario, we observe that the results differ depending on the scenario. In Scenario 1, M2 outperforms M1 and M3 in terms of bias and MSE, while the estimates obtained with M2 had a greater variance than the estimates obtained with M1 or M3. On the other hand, in Scenario 2, there are no differences among the results obtained with the three methods. The results also show that the method M2 performs correctly in both scenarios and the results are quite similar in terms of bias (which is negligible in all scenarios) and MSE. However, the performance of M1 and M3 methods in terms of bias (and consequently, also in terms of MSE) differ depending on the scenario, being much lower in Scenario 2 than in Scenario 1. We proceed below to analyse the graphical and numerical results related to each scenario.

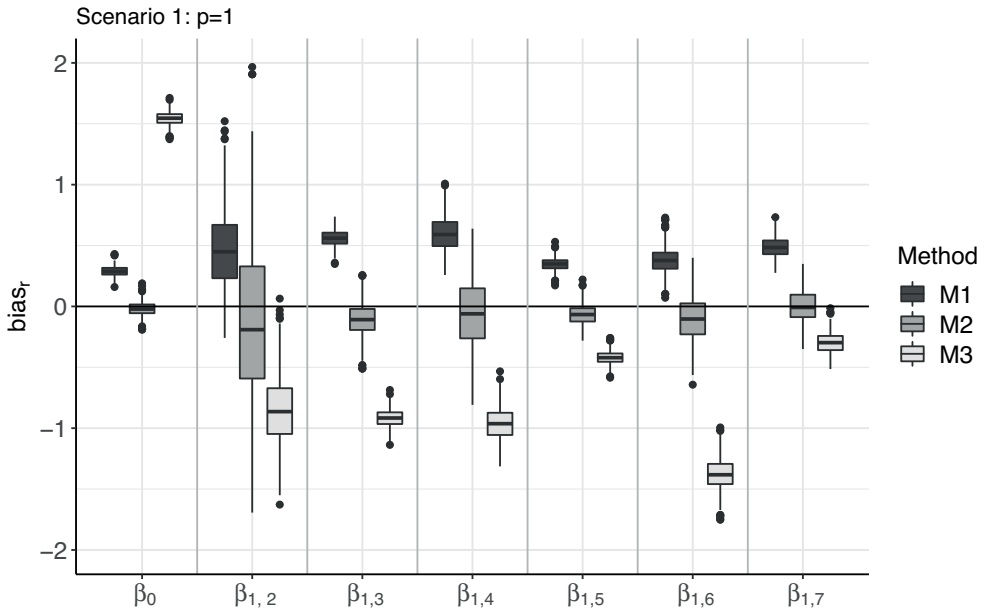
Figure 1 depicts the box-plots of the bias of the estimates obtained by the methods M1, M2 and M3 for the models with $p = 1$ (Figure 1(a)) and $p = 3$ (Figure 1(b)) covariates in Scenario 1. As can be observed, M2 is the method that performs the best in terms of bias in both models, with either $p = 1$ or $p = 3$ covariates. This can also be observed in Table 1. This table describes the numerical results of the mean, standard deviation, average bias and MSE of those estimates, as well as the true finite population coefficients in Scenario 1 for the models with $p = 1$ and $p = 3$ covariates, respectively. As can be seen, while the estimates obtained by M2 method are quite similar to the true coefficients ($\boldsymbol{\beta}^{True}$) obtained in the finite population (which leads to low average biases for this method), the estimates obtained by M1 and M3 methods differ considerably. In the estimates obtained for the model with $p = 1$ for example, for the coefficient $\beta_{1,6}$ for instance, the average bias obtained by means of M2 method is of -0.095, which is considerably lower than the one of the M1 method (0.378) and the M3 method (-1.379). It can also be observed that the average bias decreases for all the methods (and most notably for M1 and M3) when $p = 3$ covariates are included into the model. This is in line

with Nathan and Holt (1980). In particular, the average bias of the coefficient estimates related to the category $\beta_{1,6}$ decreases to 0.050 for the M1 method, to 0.007 for M2 and to -0.700 for M3 in the model with $p = 3$ covariates.

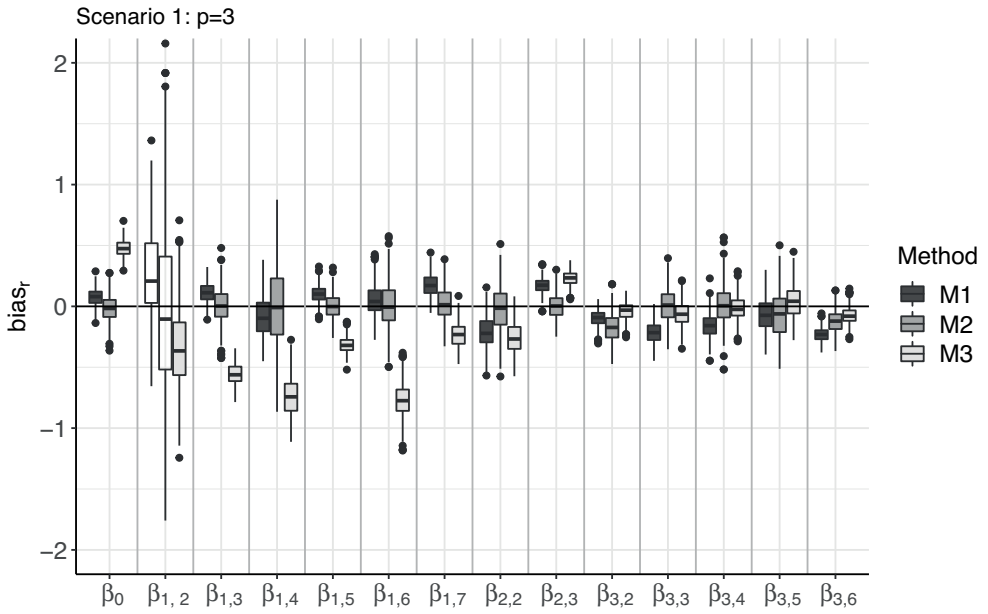
In Figure 1 it can also be seen that the variability of the estimates obtained based on the method M2 is the greatest one, comparing to the rest of the methods. This is also shown in Table 1, where the standard deviations of these estimates can be up to twice as large as that of M1 and M3. For example, the standard deviations corresponding to the estimates of $\beta_{1,3}$ are 0.063, 0.132 and 0.070 for M1, M2 and M3, respectively. The larger variability of the weighted estimates has also been observed in previous studies (see, for example, Scott and Wild (1986)). The source of variability could also be related to data. It is especially remarkable the variability of the estimates of the coefficient $\beta_{1,2}$ for all the methods in general, and most importantly for M2. It should be noted that there are very few units in the category 2 of the covariate X_1 in Scenario 1. In particular, 450 units in the population (0.2% of the total of units in the finite population) take this category on that covariate and in the samples this amount varies from 27 (0.3%) to 53 (0.7%) (results not shown). This may be affecting in the estimates of the parameter $\beta_{1,2}$, specifically for the M2 method. The behaviour of the estimates of $\beta_{1,4}$ could be explained in the same way, for which a greater variability is also observed, especially for M2 (2008 units (1.0%) in the finite population, from 178 (2.5%) to 232 (3.2%) in the samples). In addition, in Table 1, it should also be noted that for all the methods, the standard deviation of the three methods are slightly greater for the model with $p = 3$ than for the one with $p = 1$ covariates.

Finally, as shown in Table 1, the method M3 is, in most of the cases, the one with the greatest MSE, because of the large bias of the estimates based on that method. For instance, the MSE of the coefficient corresponding to the category $\beta_{1,4}$ in the model with $p = 1$ is 0.722 for the method M3, while for the M2 and M1 methods the MSE are 0.085 and 0.378, respectively. Given that the bias decreases while adding covariates for the methods M1 and M3, the MSE also decreases in the same way. For the same coefficient, when $p = 3$, the MSE related to the method M3 decreases to 0.536. The MSE of the M2 method is quite similar in both models, with $p = 1$ and $p = 3$ covariates. Comparing the MSE of M2 and M1 methods it can be observed that the MSE of M1 is greater when $p = 1$. However, in Scenario 1 with $p = 3$, there are no differences in terms of MSE between M1 and M2 due to the larger variability of M2 estimates despite their smaller bias.

Figure 2 depicts the box-plots of the bias of the estimates obtained by the methods M1, M2 and M3 for the models with $p = 1$ and $p = 4$ covariates in Scenario 2. In this case, as shown in Figure 2, the performance of the three methods is quite similar in terms of bias and variability. The differences are not considerable, neither among the different methods, nor between the different models (fitted with $p = 1$ and $p = 4$ covariates). Table 2 describes the numerical results of the mean, standard deviation, average bias and MSE of those estimates and the true finite population coefficients for $p = 1$ and $p = 4$ in Scenario 2. The average bias is very low for all the methods and in both models, either

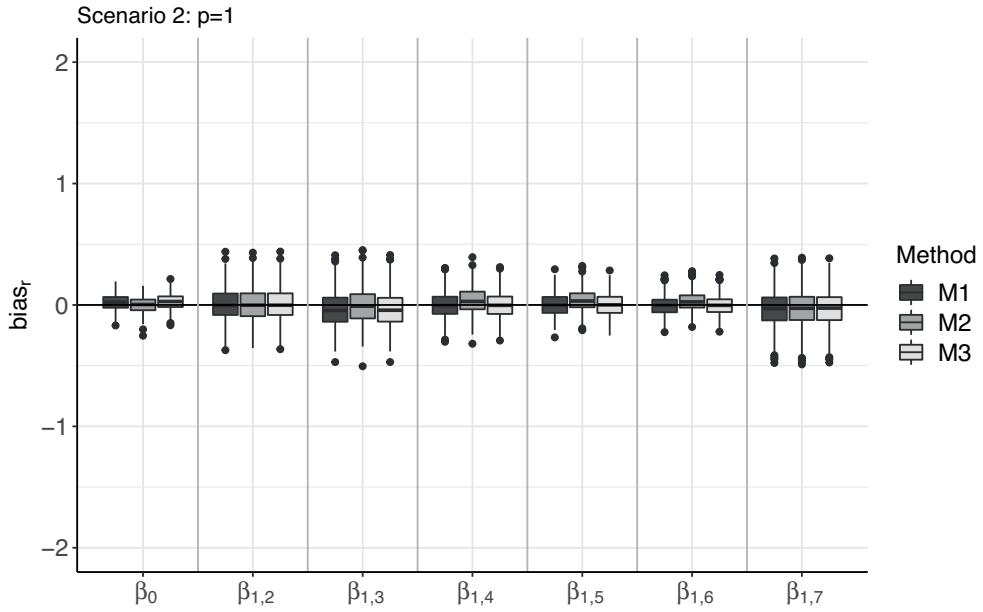


(a)

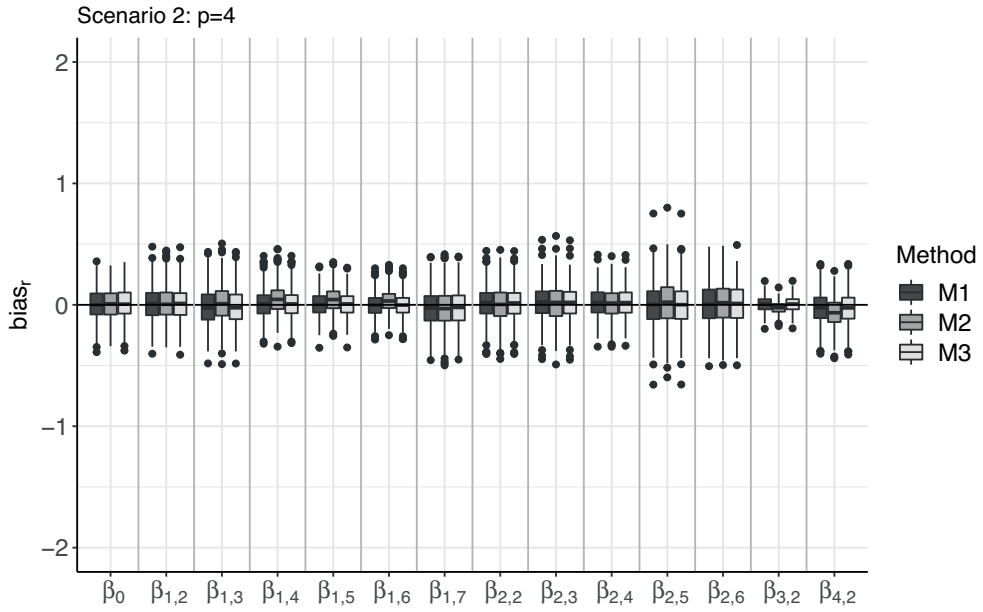


(b)

Figure 1. *Box-plots of the bias of the estimates obtained by the methods M1, M2 and M3 for the coefficients in the models with (a)* $p = 1\,(l = 7)$ *and (b)* $p = 3\,(l = 14)$ *covariates in Scenario 1,* $\forall r = 1, \ldots, R$.



(a)



(b)

Figure 2. *Box-plots of the bias of the estimates obtained by the methods M1, M2 and M3 for the coefficients in the models with (a)* $p = 1$ $(l = 7)$ *and (b)* $p = 4$ $(l = 14)$ *covariates in Scenario 2,* $\forall r = 1, \ldots, R$.

Table 1. *True finite population model coefficients ($\boldsymbol{\beta}^{True}$) and the average (mean), standard deviation (sd), average bias (AvBias) and MSE of the estimates obtained by the M1, M2 and M3 methods for the models with $p = 1$ and $p = 3$ covariates in Scenario 1 for $R = 500$ samples.*

	$\boldsymbol{\beta}^{True}$	M1			M2			M3		
		Mean (sd)	AvBias	MSE	Mean (sd)	AvBias	MSE	Mean (sd)	AvBias	MSE
p = 1 (l = 7)										
β_0	-1.015	-0.727 (0.043)	0.288	0.085	-1.035 (0.058)	-0.021	0.004	0.529 (0.056)	1.544	2.387
$\beta_{1,2}$	1.184	1.658 (0.325)	0.474	0.330	1.050 (0.659)	-0.134	0.451	0.334 (0.290)	-0.850	0.807
$\beta_{1,3}$	1.360	1.920 (0.063)	0.560	0.318	1.252 (0.132)	-0.108	0.029	0.443 (0.070)	-0.916	0.845
$\beta_{1,4}$	1.342	1.942 (0.138)	0.600	0.378	1.283 (0.286)	-0.060	0.085	0.382 (0.141)	-0.960	0.942
$\beta_{1,5}$	0.537	0.885 (0.055)	0.348	0.124	0.471 (0.085)	-0.067	0.012	0.118 (0.053)	-0.420	0.179
$\beta_{1,6}$	1.908	2.286 (0.110)	0.378	0.155	1.813 (0.184)	-0.095	0.043	0.531 (0.129)	-1.377	1.912
$\beta_{1,7}$	0.470	0.956 (0.080)	0.487	0.243	0.471 (0.121)	0.001	0.015	0.174 (0.088)	-0.296	0.095
p = 3 (l = 14)										
β_0	-0.959	-0.883 (0.069)	0.076	0.010	-0.979 (0.111)	-0.020	0.013	-0.484 (0.066)	0.475	0.230
$\beta_{1,2}$	0.602	0.850 (0.356)	0.248	0.188	0.582 (0.668)	-0.020	0.445	0.257 (0.330)	-0.344	0.227
$\beta_{1,3}$	0.824	0.937 (0.078)	0.112	0.019	0.830 (0.146)	0.006	0.021	0.269 (0.082)	-0.556	0.316
$\beta_{1,4}$	0.926	0.840 (0.158)	-0.086	0.032	0.926 (0.310)	0.000	0.096	0.184 (0.161)	-0.742	0.576
$\beta_{1,5}$	0.382	0.482 (0.066)	0.101	0.014	0.383 (0.096)	0.002	0.009	0.065 (0.064)	-0.317	0.104
$\beta_{1,6}$	1.145	1.195 (0.124)	0.050	0.018	1.153 (0.197)	0.007	0.039	0.374 (0.138)	-0.771	0.614
$\beta_{1,7}$	0.355	0.526 (0.089)	0.172	0.037	0.373 (0.125)	0.018	0.016	0.119 (0.097)	-0.235	0.065
$\beta_{2,2}$	-0.630	-0.843 (0.136)	-0.212	0.064	-0.647 (0.182)	-0.016	0.033	-0.892 (0.121)	-0.262	0.083
$\beta_{2,3}$	0.036	0.207 (0.058)	0.170	0.032	0.040 (0.098)	0.004	0.010	0.266 (0.056)	0.230	0.056
$\beta_{3,2}$	0.042	-0.055 (0.065)	-0.097	0.014	-0.128 (0.115)	-0.171	0.042	0.005 (0.067)	-0.038	0.006
$\beta_{3,3}$	1.514	1.297 (0.086)	-0.217	0.054	1.520 (0.143)	0.006	0.020	1.448 (0.091)	-0.066	0.013
$\beta_{3,4}$	1.418	1.260 (0.096)	-0.158	0.034	1.428 (0.150)	0.010	0.023	1.402 (0.098)	-0.016	0.010
$\beta_{3,5}$	1.564	1.491 (0.134)	-0.073	0.023	1.497 (0.190)	-0.067	0.040	1.601 (0.133)	0.037	0.019
$\beta_{3,6}$	2.700	2.465 (0.053)	-0.234	0.058	2.577 (0.093)	-0.123	0.024	2.623 (0.069)	-0.077	0.011

Table 2. *True finite population model coefficients ($\boldsymbol{\beta}^{True}$) and the average (mean), standard deviation (sd), average bias (AvBias) and MSE of the estimates obtained by the M1, M2 and M3 methods for the models with $p = 1$ and $p = 4$ covariates in Scenario 2 for $R = 500$ samples.*

	$\boldsymbol{\beta}^{True}$	M1			M2			M3		
		Mean (sd)	AvBias	MSE	Mean (sd)	AvBias	MSE	Mean (sd)	AvBias	MSE
p = 1 (l = 7)										
β_0	-0.965	-0.945 (0.068)	0.020	0.005	-0.965 (0.065)	0.000	0.004	-0.941 (0.068)	0.025	0.005
$\beta_{1,2}$	2.676	2.681 (0.127)	0.005	0.016	2.680 (0.133)	0.003	0.018	2.682(0.127)	0.006	0.016
$\beta_{1,3}$	3.369	3.337(0.149)	-0.032	0.023	3.366 (0.152)	-0.003	0.023	3.338 (0.149)	-0.031	0.023
$\beta_{1,4}$	3.539	3.539 (0.110)	0.000	0.012	3.576 (0.109)	0.038	0.013	3.539 (0.109)	0.001	0.012
$\beta_{1,5}$	3.034	3.035 (0.093)	0.002	0.009	3.071 (0.091)	0.037	0.010	3.036 (0.093)	0.003	0.009
$\beta_{1,6}$	1.571	1.568 (0.082)	-0.003	0.007	1.600 (0.075)	0.029	0.006	1.569 (0.082)	-0.002	0.007
$\beta_{1,7}$	-2.854	-2.886 (0.139)	-0.033	0.020	-2.883 (0.146)	-0.029	0.022	-2.884 (0.139)	-0.030	0.020
p = 4 (l = 14)										
β_0	-1.529	-1.523 (0.125)	0.006	0.016	-1.524 (0.127)	0.005	0.016	-1.516 (0.125)	0.012	0.016
$\beta_{1,2}$	2.478	2.490 (0.134)	0.012	0.018	2.490 (0.140)	0.012	0.020	2.490 (0.134)	0.012	0.018
$\beta_{1,3}$	3.206	3.188 (0.153)	-0.017	0.024	3.221 (0.156)	0.015	0.024	3.189 (0.153)	-0.017	0.024
$\beta_{1,4}$	3.329	3.342 (0.121)	0.013	0.015	3.379 (0.122)	0.050	0.017	3.341 (0.121)	0.013	0.015
$\beta_{1,5}$	2.780	2.785 (0.103)	0.005	0.011	2.824 (0.103)	0.044	0.012	2.786 (0.103)	0.006	0.011
$\beta_{1,6}$	1.379	1.376 (0.097)	-0.003	0.009	1.413 (0.092)	0.034	0.010	1.378 (0.097)	-0.001	0.009
$\beta_{1,7}$	-2.943	-2.974 (0.149)	-0.031	0.023	-2.974 (0.157)	-0.031	0.026	-2.970 (0.150)	-0.028	0.023
$\beta_{2,2}$	0.702	0.715 (0.134)	0.013	0.018	0.704 (0.144)	0.002	0.021	0.714 (0.134)	0.012	0.018
$\beta_{2,3}$	1.391	1.407 (0.140)	0.016	0.020	1.404 (0.148)	0.013	0.022	1.406 (0.140)	0.015	0.020
$\beta_{2,4}$	0.811	0.827 (0.117)	0.016	0.014	0.819 (0.126)	0.008	0.016	0.827 (0.117)	0.016	0.014
$\beta_{2,5}$	1.620	1.618 (0.174)	-0.001	0.030	1.638 (0.185)	0.018	0.035	1.618 (0.174)	-0.001	0.030
$\beta_{2,6}$	1.664	1.669 (0.159)	0.005	0.025	1.678 (0.165)	0.014	0.027	1.670 (0.159)	0.006	0.025
$\beta_{3,2}$	-0.427	-0.422 (0.057)	0.005	0.003	-0.451 (0.049)	-0.024	0.003	-0.421 (0.057)	0.005	0.003
$\beta_{4,2}$	-0.481	-0.506 (0.129)	-0.025	0.017	-0.548 (0.123)	-0.067	0.019	-0.508 (0.129)	-0.027	0.017

with $p = 1$ or $p = 4$ covariates. The greatest observed average bias is -0.067, which corresponds to the coefficient $\beta_{4,2}$ of the model with $p = 4$ covariates for the method M2. The variability of the estimates obtained by the method M2 are usually slightly greater than that of the rest of the methods. However, as noted above, those differences are very small. The greatest difference in terms of standard deviation of the estimates and MSE are observed in the model with $p = 4$ for the coefficient estimates corresponding to category $\beta_{2,5}$. The standard deviation of the estimates obtained by means of M2 method is 0.185 while the ones corresponding to the M1 and M3 methods are 0.174. In the same way, the MSE of the M2 method for this coefficient is 0.035 while for the methods M1 and M3 is 0.030. It can be concluded that all the studied methods perform properly to estimate the finite population model coefficients in Scenario 2.

5. Application to the real data sets

In this section we apply the methods described in Section 3 to the real survey data described in Section 2. The goal is to compare the coefficient estimates obtained by means of the different methods among them. Note that in this case, the real finite population coefficients are not known.

Table 3. *Coefficient estimates (Estimate) and their standard errors (SE) obtained by means of the methods M1, M2 and M3 for the ESIE survey with $p = 3$ covariates.*

	ESIE survey					
	M1		M2		M3	
	Estimate	SE	Estimate	SE	Estimate	SE
β_0	-2.261	0.097	-2.482	0.133	-2.217	0.140
$\beta_{1,2}$	1.892	0.338	1.293	0.444	1.697	0.368
$\beta_{1,3}$	2.490	0.107	2.718	0.161	2.337	0.119
$\beta_{1,4}$	2.248	0.196	2.577	0.299	2.151	0.215
$\beta_{1,5}$	1.550	0.084	1.721	0.111	1.458	0.094
$\beta_{1,6}$	2.260	0.146	2.544	0.206	2.092	0.181
$\beta_{1,7}$	1.341	0.103	1.130	0.133	1.197	0.119
$\beta_{2,2}$	-0.774	0.148	-0.613	0.189	-0.883	0.329
$\beta_{2,3}$	0.453	0.073	0.358	0.107	0.538	0.123
$\beta_{3,2}$	0.669	0.069	0.632	0.097	0.750	0.077
$\beta_{3,3}$	0.996	0.096	0.965	0.132	1.124	0.134
$\beta_{3,4}$	1.479	0.114	1.452	0.152	1.698	0.149
$\beta_{3,5}$	2.230	0.182	2.205	0.241	2.461	0.209
$\beta_{3,6}$	2.454	0.143	2.532	0.151	2.787	0.195

One model was fitted to each of the surveys. In particular, we fitted the model with three covariates ($p = 3$) to the ESIE survey and the model with four covariates ($p = 4$)

to the PRA survey. Those covariates are the ones that were considered in the simulation study for both surveys and are also considered in the models that are applied in practice by Eustat. To fit those models, the three methods described in Section 3 were applied: the unweighted logistic regression (M1), the weighted logistic regression (M2) and the unweighted logistic regression with random intercept (M3). Table 3 and Table 4 depict the coefficient estimates and their standard errors obtained for models fitted to the ESIE and PRA surveys respectively.

Table 4. *Coefficient estimates (Estimate) and their standard errors (SE) obtained by means of the methods M1, M2 and M3 for the PRA survey with $p = 4$ covariates.*

	PRA survey					
	M1		M2		M3	
	Estimate	SE	Estimate	SE	Estimate	SE
β_0	-2.039	0.176	-2.040	0.171	-2.037	0.179
$\beta_{1,2}$	2.508	0.164	2.523	0.172	2.515	0.164
$\beta_{1,3}$	3.106	0.179	3.105	0.191	3.113	0.179
$\beta_{1,4}$	3.191	0.121	3.292	0.126	3.194	0.122
$\beta_{1,5}$	2.836	0.114	2.934	0.118	2.835	0.114
$\beta_{1,6}$	1.455	0.103	1.543	0.108	1.454	0.103
$\beta_{1,7}$	-3.170	0.184	-3.102	0.199	-3.182	0.184
$\beta_{2,2}$	1.005	0.174	0.899	0.177	1.016	0.174
$\beta_{2,3}$	1.689	0.178	1.587	0.182	1.700	0.179
$\beta_{2,4}$	1.167	0.171	1.056	0.170	1.170	0.172
$\beta_{2,5}$	2.123	0.207	1.970	0.227	2.128	0.208
$\beta_{2,6}$	2.357	0.192	2.177	0.201	2.360	0.193
$\beta_{3,2}$	-0.596	0.063	-0.546	0.067	-0.596	0.063
$\beta_{4,2}$	0.547	0.158	0.530	0.190	0.551	0.159

As shown in Table 3, the coefficient estimates, as well as their standard errors, obtained by means of the three above-mentioned methods differ considerably in the ESIE survey. It should be noted that these differences in the estimations and their standard errors, could lead to considerable differences in the Wald statistic defined as the fraction among those parameters. However, in this case, those differences did not affect the significance of the model parameters and all of them are statistically significant (results not shown). The largest standard errors are in most of the cases the ones obtained by means of the method M2. In addition, the standard errors related to the coefficient $\beta_{1,2}$ are larger than any other's, which is in line with the large variability observed in the simulation study for this coefficient (in Scenario 1). Based on the results obtained in the simulation study, we may conclude that the model fitted by the method M2 would be the preferred one in this case.

In contrast, the coefficient estimates and their standard errors obtained for the PRA survey are very similar among them, as can be observed in Table 4. This is also in line with the results observed in the simulation study (in Scenario 2). As expected, the standard errors of the estimates obtained by M2 are usually slightly greater than the rest, although there are not great differences, in general.

6. Discussion

In this work we compared the performance of three different methods to estimate model coefficients in the logistic regression framework for complex survey data by means of a real data based simulation study. In general, the results we obtained are in line with the ones obtained in related works, based on either logistic (Scott and Wild, 1986, 2002; Lumley and Scott, 2017; Chambless and Boyle, 1985; Reiter et al., 2005) or linear regression framework (DeMets and Halperin, 1977; Holt et al., 1980; Nathan and Holt, 1980; Smith, 1981). Nevertheless, there are also some differences among this work and the above-mentioned studies. We proceed to comment on these similarities and differences in the following lines.

One of the greatest differences between this study and the ones mentioned previously is that this work is a simulation study based on real survey data. The objective has been to work in a realistic scenario that allows us to compare the results we obtain to the true coefficients of the finite population models. Data for the simulation study have been simulated based on two real surveys conducted by the Official Statistics Basque Office (Eustat). In both surveys the finite population were sampled by single-stage stratification. However, the strata were defined in very different ways. In the ESIE survey the strata were defined by means of the combination of three categorical variables with many categories, resulting in a total of 585 small strata. On the other hand, in the PRA survey, strata were defined by means of the region to which each individual belongs, which leads to 23 different strata. In addition to the sampling design, the impact of the number of covariates included in the model and the number of parameters, were also analysed. It should be noted that in this simulation study the theoretical model from which the finite population is generated from is not known for us. Thus, we compare the model estimates obtained based on the methods under study to the true coefficient values obtained by fitting the model to the finite population.

The main conclusions of this study are that the weighted logistic regression (M2) performed properly in both scenarios and the estimates we obtained were unbiased. In contrast, the behavior of the unweighted logistic regression (M1) and the unweighted logistic regression with random intercept (M3) depended on the scenario and on the number of covariates/parameters estimated in the model. In the scenario related to the ESIE survey, unlike in the scenario based on the PRA survey, biased estimates were obtained based on these two methods. These results are in line with Scott and Wild (1986); Holt et al. (1980); Nathan and Holt (1980) among others, which also warn about the bias of the unweighted coefficient estimates in both, linear and logistic regression

frameworks. Scott and Wild (1986) claim that the bias of the unweighted coefficient estimates is smaller when the model fitted to the sample is exactly the same as the true theoretical model from which the data is derived than when the model fitted is "reasonable but not perfect". As mentioned previously, the theoretical model from which the finite population is generated from is not known for us. Nevertheless, in this study we have also observed that the bias becomes smaller when more covariates are included into the model, which would be in line with the results obtained in the above-mentioned studies. However, this bias is still larger than the bias obtained by means of the weighted logistic regression. For this reason, the message we aim to transmit with this work is the recommendation of fitting weighted models. In line with Reiter et al. (2005), we agree that comparing the estimates obtained with the unweighted model can help to detect if the model is well specified (and improve the model, if the needed variables are available), since a large difference between the two estimates can suggest that the fitted model is mis-specified.

The variability of the estimates obtained by the weighted logistic regression model is greater than that of the estimates obtained by means of the unweighted logistic regression model (with and/or without random intercept) which is in line with Chambless and Boyle (1985); Lumley and Scott (2017); Scott and Wild (1986). These differences are not very large in most of the cases. However, we have observed that when there are few individuals in a particular category of a categorical variable, then the variability of the weighted estimates of the coefficient corresponding to that category can be much greater than the unweighted ones. We conclude that we should be careful when we have categorical variables with unbalanced distribution of individuals in the categories. In addition, we should keep in mind that cluster sampling has not been considered in this simulation study and that estimates with sampling weights may show higher variability in this context. In addition, the mixed model is commonly used when we have clustered sampling, being the clusters the ones used as random effects, instead of the strata, as in this paper. It should be noted that in the simulation study we have conducted, it was unfeasible to put all the design information as fixed effect (as recommended for strata) because of the problems that would arise for both model estimation and interpretation. For this reason, we have opted to use the strata as a random effect. Through this study we have been able to verify that the mixed model does not provide us with advantages compared to the other models, but, in order to make a really fair comparison for these models we should also test what happens when we have a design that involves clustering (further research).

We also applied the three methods under study to real survey data and the estimates we obtained are in line with the results observed in the simulation study. On the one hand, in the PRA survey, the estimates are quite similar among them, and there are not many differences between the standard deviations of these estimates, which leads us to conclude that all the studied methods work properly in this case. On the other hand, in the ESIE survey, there are many differences in the estimates of the parameters among different methods. Observing the similarities among the simulation study and the

application to real data sets, and taking into account that those results are also in line with the results obtained in similar empirical studies, such as Chambless and Boyle (1985) and Lumley and Scott (2017), we can assume that the weighted logistic regression would be preferred when working with ESIE survey data.

We now proceed to comment on the limitations of this work. First of all, in this simulation study we are unable to know which is the theoretical model from which the data is derived due to the fact that we aimed for the simulation study to be based on real survey data and hence, we have focused on comparing the estimates obtained based on the samples with the true coefficients of the model fitted to the finite population. It should be noted that often the objective in working with survey data is to draw conclusions related to that particular finite population, and therefore, this comparative study makes sense in that context. For those readers who are interested in comparisons with the theoretical infinite population model, we suggest checking Scott and Wild (2002). Secondly, as mentioned above, some authors recommend including the design variables and the interactions between them as covariates in the model. However, in this case, and in particular in the case of the ESIE survey, this option would not be feasible due to the large number of parameters (a total of 585) to be estimated within the model. Therefore, we have decided to fit the mixed model, replicating in this way the comparison made by Lumley and Scott (2017) on real datasets (it should be noted that we considered strata as random effect instead of clusters, as in the referenced paper). In addition, some of the covariates included in the models are related to the stratification variables. It should also be noted that in this simulation study we have worked with surveys of considerable sample sizes, which is quite common in official statistics. Nevertheless, we also believe that it would be interesting to work with simulations based on real surveys with smaller sample sizes and compare the results, paying special attention to the variability of the estimates. It should also be noted as a limitation, the fact that in this work we consider 100% of response, ignoring in this way the impact that non-response may have on the sampling weights, which is a common problem to deal with in the daily practice of complex survey data. Lastly, in this work we have focused on the estimation of the parameters of the logistic regression model. Other issues of interest, such as the selection of the covariates or the effect that these differences may have on the estimated probabilities of the individuals, are out of the scope of this work.

To sum up, the weighted logistic regression performs properly in all the scenarios we have drawn. In contrast, the behavior of the unweighted logistic regression (both, with and without random intercept) depends on the scenario. Therefore, based on the results of the simulation study, we believe that not using sampling weights when necessary leads to worse results than using them when they are not needed. For this reason, we would recommend the use of the weighted logistic regression model in the context of complex survey data.

Nevertheless, we are aware that the use of sampling weights is an ongoing debate. For instance, Lumley (2010) pointed out two points why it might be interesting not to use weights in certain cases. One of the reasons was the lack of software available

to work with them, something that is nowadays solved as they note. Another reason why some researchers may find it interesting to ignore the weights in cases where the estimates are similar is that greater standard errors are obtained with weighted estimates, involving reduction of precision. However, as the same authors point out, small biases of unweighted estimates cannot be reliably detected from real data and could be enough to give less-accurate estimates than the weighted methods. It is possible that the resistance of some researchers to use sampling weights comes from the feeling of not knowing how to work with them. For example, Gelman (2007) defined survey weighting as a "mess". We would like to end this discussion with a comment on our view of the importance of using sampling weights in the development of a prediction model as a whole. When we fit prediction models, we are usually not only interested in the estimation of the model itself, but we are usually interested in developing good prediction models that can be used in daily practice. As Steyerberg and Vergouwe (2014) pointed out, several steps should be considered to develop valid prediction models that could be applied in practice, that go from the estimation of model parameters to the validation of the final model. Therefore, the authors would like to highlight the lack of tools to develop good prediction models as a disadvantage of using weighted techniques. Although steps are being taken in this direction as several works published over the last decade show (Lumley and Scott, 2015, Yao, Li and Graubard, 2015; Lumley, 2017; Wieczorek, Guerin and McMahon, 2022; Iparragirre et al., 2022), the authors believe that it is essential to continue research in this line, lose the fear of sampling weights and continue to improve techniques so that we can develop good prediction models considering complex sampling designs.

Acknowledgment

This work was financially supported in part by grants from the Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco [IT1456-22] and by the Ministry of Science and Innovation through BCAM Severo Ochoa accreditation [CEX2021-001142-S / MICIN / AEI / 10.13039/501100011033] and through project [PID2020-115882RB-I00 / AEI / 10.13039/501100011033] funded by Agencia Estatal de Investigación and acronym "S3M1P4R" and also by the Basque Government through the BERC 2022-2025 program. The work of AI was supported by grant [PIF18/213].

We would like to acknowledge the Official Statistics Basque Office (Eustat) for providing us with the ESIE and PRA survey data.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Binder, D. A. (1981). On the variances of asymptotically normal estimators from complex surveys. *Survey Methodology*, 7(3):157–170.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review / Revue Internationale de Statistique*, 51(3):279–292.

Binder, D. A. and Roberts, G. (2009). Design- and model-based inference for model parameters. In *Handbook of Statistics*, volume 29, pages 33–54. The Netherlands: Elsevier.

Brewer, K. R. W. and Mellor, R. W. (1973). The effect of sample structure on analytical surveys. *Australian Journal of Statistics*, 15(3):145–152.

Chambers, R. L. and Skinner, C. J. (2003). *Analysis of Survey Data*. New York: John Wiley & Sons.

Chambless, L. E. and Boyle, K. E. (1985). Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Communications in Statistics-Theory and Methods*, 14(6):1377–1392.

DeMets, D. and Halperin, M. (1977). Estimation of a simple regression coefficient in samples arising from a sub-sampling procedure. *Biometrics*, 33(1):47–56.

Diggle, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data (2nd ed.)*. Oxford: Oxford University Press.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164.

Hausman, J. A. and Wise, D. A. (1981). Stratification on endogenous variables and estimation: The Gary income maintenance experiment. In *Structural Analysis of Discrete Data with Econometric Applications*, pages 365–391. Cambridge, MA: MIT Press.

Heeringa, S. G., West, B. T., and Berglund, P. A. (2017). *Applied Survey Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Holt, D., Smith, T. M. F., and Winter, P. D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society: Series A (General)*, 143(4):474–487.

Iparragirre, A., Barrio, I., Aramendi, J., and Arostegui, I. (2022). Estimation of cut-off points under complex-sampling design data. *SORT-Statistics and Operations Research Transactions*, 46(1):137–158.

Korn, E. L. and Graubard, B. I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49(3):291–295.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4):987–1006.

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: another view. *Statistical Science*, 19(2):219–238.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Sons.

Lumley, T. (2017). Pseudo-R2 statistics under complex sampling. *Australian & New Zealand Journal of Statistics*, 59(2):187–194.

Lumley, T. (2019). survey: analysis of complex survey samples. R package version 3.35-1.

Lumley, T. and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1):1–18.

Lumley, T. and Scott, A. (2017). Fitting regression models to survey data. *Statistical Science*, 32(2):265–278.

Masood, M., Newton, T., and Reidpath, D. (2016). Comparison of four analytic strategies for complex survey data: a case-study of Spanish data. *Epidemiology, Biostatistics and Public Health*, 13(1):1–7.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (2nd ed.)*. London: Chapman & Hall.

Nathan, G. and Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(3):377–386.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. In *Handbook of Statistics*, volume 29, pages 455–487. The Netherlands: Elsevier.

Reiter, J. P., Zanutto, E. L., and Hunter, L. W. (2005). Analytical modeling in complex surveys of work practices. *ILR Review*, 59(1):82–100.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Scott, A. J. (1977). On the problem of randomization in survey sampling. *Sankhya*, 39:1–9.

Scott, A. J. and Wild, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2):170–182.

Scott, A. J. and Wild, C. J. (2002). On the robustness of weighted methods for fitting models to case–control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):207–219.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons.

Smith, T. M. F. (1981). Regression analysis for complex surveys. In *Current Topics in Survey Sampling*, pages 267–292. New York: Academic Press.

Steyerberg, E. W. and Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal*, 35(29):1925–1931.

Sugden, R. A. and Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71(3):495–506.

Wieczorek, J., Guerin, C., and McMahon, T. (2022). K-fold cross-validation for complex sample surveys. *Stat*, 11(1):e454.

Wolter, K. (2007). *Introduction to Variance Estimation (2nd ed.)*. New York: Springer-Verlag.

Yao, W., Li, Z., and Graubard, B. I. (2015). Estimation of ROC curve with complex survey data. *Statistics in medicine*, 34(8):1293–1303.

Appendices

A. Pseudo-population generation

This section describes the process of generating the pseudo-populations that have been used in the simulation study described in Section 4, in Scenario 1 (based on the ESIE survey) and in Scenario 2 (based on the PRA survey).

The pseudo-population applied in Scenario 2, related to the PRA survey, is actually a real finite population, for which the response variable, as well as the rest of the explanatory variables, are known. This pseudo-population was obtained and provided by Eustat.

In the case of Scenario 1, we have generated a pseudo-population based on the real finite population and sample of the ESIE survey. Let us denote as S_{ESIE} the original survey sample and U_{ESIE} the real finite population of size N ($S_{\text{ESIE}} \subset U_{\text{ESIE}}$). As explained in Section 2, a total of H strata have been defined (i.e., $\{1,\ldots,H\}$) combining information of three categorical variables, which will be denoted as X_1, X_2 and X_3. Therefore, the finite population can be partitioned in subsets defined by means of these strata, i.e., $U_{\text{ESIE}} = \bigcup_{h=1}^{H} U_{\text{ESIE},h}$. $\forall h \in \{1,\ldots,H\}$ let us indicate as N_h the size of stratum h in the finite population U_{ESIE} ($U_{\text{ESIE},h}$) and as n_h the size of this stratum in the sample S_{ESIE}. Then, the sampling weight associated to a unit $j \in S_{\text{ESIE}}$ from stratum h is the following:

$$w_j = \frac{N_h}{n_h}. \tag{12}$$

Our goal is to generate a pseudo-population (U) based on the known real ESIE survey data, for which all the information of the covariates $X_1,\ldots,X_p$ and the response variables $Y_1,\ldots,Y_q$ will be available. This new pseudo-population U will be the same size as the true ESIE population (N). In order to ease the notation, the variable names of the pseudo-population are the same as in the real finite population and the units of the real ESIE population will be denoted as $j \in U_{\text{ESIE}}$ while the units that are artificially generated for the pseudo-population will be denoted as $i \in U$.

Several dichotomous response variables are available in the original survey being the response variable Y, the one we have applied in the simulation study, one of them. All possible combinations of these response variables have been examined. For instance, assuming that $Y_1,\ldots,Y_q$ are all the response variables that are available in the survey (where $Y \in \{Y_1,\ldots,Y_q\}$), for some $j \in S_{\text{ESIE}}$: $\boldsymbol{y}_j = (y_{1,j},\ldots,y_{q,j}) = \alpha$, $\forall \alpha \in \{\alpha_1,\ldots,\alpha_A\}$, where $\{\alpha_1,\ldots,\alpha_A\}$ is the set of all of possible combinations of the responses. For

each stratum $h \in \{1,\ldots,H\}$ and for each possible combination of the responses (i.e., $\forall \alpha \in \{\alpha_1,\ldots,\alpha_A\}$) we generate $N_{h,\alpha}$ units in the pseudo-population (U) as:

$$N_{h,\alpha} = \sum_{j \in S_{\text{ESIE}}} w_j 1_{U_{\text{ESIE},h}}(j)\,[\boldsymbol{y}_j = \alpha], \tag{13}$$

where,

$$1_{U_{\text{ESIE},h}}(j) = \begin{cases} 1, & \text{if } j \in U_{\text{ESIE},h}, \\ 0, & \text{if } j \notin U_{\text{ESIE},h}, \end{cases} \tag{14}$$

and

$$[\boldsymbol{y}_j = \alpha] = \begin{cases} 1 & \text{if} \quad \boldsymbol{y}_j = \alpha, \\ 0 & \text{if} \quad \boldsymbol{y}_j \neq \alpha. \end{cases} \tag{15}$$

In this way, $N_{h,\alpha}$ is the number of units of the pseudo-population U in stratum h, which take the values of responses $(y_{1,j},\ldots,y_{q,j}) = \alpha$. Once we repeat the process for $\forall h \in \{1,\ldots,H\}$ and $\forall \alpha \in \{\alpha_1,\ldots,\alpha_A\}$ a pseudo-population of $N = \sum_{h\in\{1,\ldots,H\}} \sum_{\alpha\in\{\alpha_1,\ldots,\alpha_A\}} N_{h,\alpha} = \sum_{j\in S_{\text{ESIE}}} w_j$ units we generated with the information of response variables (Y, among others) and strata (hence, information fo the design variables X_1, X_2 and X_3 will also be generated).

Finally we generate the rest of the ovariates as follows $\forall d \in \{4,\ldots,p\}$ assume that X_d is a categorical variable with a total of K categories: $\{1,\ldots,K\}$ categories. Then, for each unit i generated in the pseudo-population ($\forall i \in U$), we generated $x_{di} \in \{1,\ldots,K\}$ following a categorical distribution (i.e., $x_{di} \sim Cat(\pi_{d1},\ldots,\pi_{dK})$) where the probability corresponding to each category $k \in \{1,\ldots,K\}$ is calculated as follows based on the known ESIE finite population U_{ESIE}. Let us assume that $i \in U_{ESIE,h}, \forall h \in \{1,\ldots,H\}$. Then,

$$\pi_{dk} = \frac{\sum_{j\in U_{\text{ESIE}}} 1_{U_{\text{ESIE},h}}(j)\,[x_{dj} = k]}{\sum_{j\in U_{\text{ESIE}}} 1_h(j)}, \forall k \in \{1,\ldots,K\}, \tag{16}$$

where $1_{U_{\text{ESIE},h}}(j)$ is defined in (14) and,

$$[x_{dj} = k] = \begin{cases} 1 & \text{if} \quad x_{dj} = k, \\ 0 & \text{if} \quad x_{dj} \neq k, \end{cases} \qquad \forall j \in U_{\text{esie}} \text{ and } \forall k \in \{1,\ldots,K\}. \tag{17}$$

In this way, the pseudo-population based on the ESIE survey has been generated with the response variable Y, the vector of explanatory variables $\boldsymbol{X}$ and the strata.

B. Pseudo-population sampling process

The two pseudo-populations have been sampled by single-stage stratified sampling, in the same way as the real survey data described in Section 2.

In order to sample the pseudo-population of the Scenario 1, first, we identify how many units have been sampled from a stratum h, $\forall h \in \{1,\ldots,H\}$ in the real survey sample S_{ESIE} (let us denote this amount as n_h). Then, we sample n_h units randomly

from stratum h from the pseudo-population U. Repeating the same process for $\forall h \in \{1,\ldots,H\}$ we obtain a sample S.

Finally, sampling weights are assigned to each sampled unit as follows. For $\forall i \in S$ assume that i is a unit from stratum h, $\forall h \in \{1,\ldots,H\}$, then:

$$w_i = \frac{N_h}{n_h}, \tag{18}$$

where N_h indicates the number of units in the stratum h in U, and n_h the number of units in the stratum h in S.

SORT 48 (1) January-June 2024, 93-124 DOI: 10.57645/20.8080.02.15

Kernel Weighting for blending probability and non-probability survey samples

María del Mar Rueda[1,*], Beatriz Cobo[2], Jorge Luis Rueda-Sánchez[1], Ramón Ferri-García[1] and Luis Castro-Martín[2]

Abstract

In this paper we review some methods proposed in the literature for combining a non-probability and a probability sample with the purpose of obtaining an estimator with a smaller bias and standard error than the estimators that can be obtained using only the probability sample. We propose a new methodology based on the kernel weighting method. We discuss the properties of the new estimator when there is only selection bias and when there are both coverage and selection biases. We perform an extensive simulation study to better understand the behaviour of the proposed estimator.

MSC: *62D05.*

Keywords: *Kernel weighting, survey sampling, non-probability sample, coverage bias, selection bias.*

1. Introduction

Probability sampling methods are well established by statistical offices and researchers as one of the primary tools for data collection in surveys. This is because when controlling the sampling design, it is feasible to make valid statistical inference about large finite populations using relative small samples. There exists an extensive literature on methods for probability sampling and design-based inferences for complex surveys.

[*] *Corresponding author:* mrueda@ugr.es

[1] Department of Statistics and Operational Research, University of Granada, Spain.

[2] Department of Quantitative Methods for Economics and Business, University of Granada, Spain.

Received: November 2022
Accepted: September 2023

However, the deployment of probability sampling methods has become more challenging, as there has been a notorious decline in response rates (Marken, 2018; Kennedy and Hartig, 2019) with the subsequent increase of the survey costs. In addition, new data sources which have arisen in recent years could be considered as alternatives to survey data. Examples are large volume datasets coming from sources such as passive data or "data lakes", and web surveys that have the potential of providing more timely estimates, as well as offering easier data access and lower data collection costs than traditional probability sampling, leading to larger sample sizes. On the other hand, there are serious issues concerning the use of non-probability survey samples (or volunteer samples) for estimation. Non-probability surveys are those where the inclusion probability is not known and/or not strictly positive for any individual in the population, which is the case for volunteer samples obtained from the Internet or similar means. For this reason, non-probability surveys are often known as voluntary surveys. The primary issue with these data sources is that the selection mechanism, which decides what individuals are eventually included in the dataset, is often unknown and may induce serious coverage and selection biases. Coverage bias can be defined as the bias that arise from the lack of exhaustiveness of the sampling frame from which the sample is drawn, this is, the inability of the sampling frame to include all the members of the target population. Selection bias is a term that comprises different types of errors when drawing the sample, but the most common in the aforementioned data sources is self-selection: the decision of being in the sample or not is taken by the individuals themselves, meaning that the inclusion probabilities are not given by the sampling design but by the participants, and therefore these probabilities remain unknown, constituting a non-probability sample. The generalization of the results under these biases is therefore compromised.

Despite these limitations, non-probability survey designs may be particularly useful in several cases. For example, they can be used in those cases where the target population is a small subpopulation unlikely to meet sample size requirements, or when we are interested in non-demographical strata which cannot be considered in a sampling design. Given the potential of non-probability surveys, statisticians have studied the integration or combination of data from probability and non-probability samples. Some reviews on methods of statistical data integration for finite population inference can be consulted in Buelens, Burger and vanden Brakel (2018), Valliant (2020), Yang and Kim (2020) or Rao (2020). The number of papers that are emerging in recent years in this field is immense. The importance that the topic is taking has motivated the holding of specialized sessions in many statistics and survey congresses as well as a special discussion paper in the Survey Methodology journal (vol 48, n.2). The paper of Wu (2022) ably and usefully summarizes the state of the literature of analysis of non-probability survey data and comments to the article by Bailey (2022), Elliott (2022), Lohr (2022), Meng (2022) and Wang and Kim (2022) deal with aspects of great interest and topicality in this subject.

Different data integration methods, which are based on combining probability and non-probability samples, have been developed in the literature on survey sampling. These integration methods can be divided into three groups depending on the availabil-

ity of the study variable: available in the non-probability sample only, in the probability sample only, or in both samples.

Many methods consider the first case, where the target variable has been observed in the non-probability sample only. In this situation, the probability sample plays an important role as the reference data, and can be used to increase the efficiency of the estimates through a variety of adjustment approaches to account for the selection bias of non-probability samples. However, other methods were also developed from different perspectives according to the availability of auxiliary information. Calibration (Deville and Särndal, 1992; Ferri-García and Rueda, 2018), propensity score adjustment (PSA) (Lee, 2006; Lee and Valliant, 2009; Castro, Rueda and Ferri-García, 2022), kernel weighting (KW) (Wang et al., 2020), statistical matching (or mass imputation) (Rivers, 2007; Beaumont, 2020), double robust estimation (Kim and Wang, 2019) and superpopulation modeling (Valliant, Dorfman and Royall, 2000; Buelens et al., 2018) are relevant techniques to mitigate selection bias.

When the non-probability (or volunteer) survey contains auxiliary variables but no study variable, Medous et al. (2022) shows how the use of a non-probability database can improve estimates from a probability sample and they define a class of QR predictors (Särndal and Wright, 1984) asymptotically design-unbiased under certain conditions.

In this paper we consider the third situation posed above, where the study variables are measured in both samples. In Section 2 we review the estimation from probability and non-probability samples to introduce the notation and the framework. In Section 3 we revisit some important works in data integration for handling selection bias in our context. In Section 4 we adapt the kernel weighting method introduced in Wang et al. (2020), to data integration. First, we consider a situation where there are no coverage biases (there is a one-to-one correspondence between the target population and the sampling frames), and we propose a KW estimator by a linear combination of biased and unbiased estimators of a population mean. When undercoverage occurs in the non-probability sample (the sampling frame does not include all members of the target population), as is usual in practice, we propose a KW estimator based on dual frame methodology. We derive conditions such that these proposed estimators are asymptotically design-unbiased. In Section 5, we use Monte Carlo simulations to compare the proposed method with several models and show that the kernel weighted estimator is a good compromise for several setups. Finally we conclude and give perspectives in 6.

2. Context and notation

Let U be the target population of size N, $U = \{1,\ldots,i,\ldots,N\}$. Let s_v be the set of n_v units selected from the frame U_v using a non-probability (volunteer) data collection method. Let s_r be a probability sample of size n_r selected from a frame U_r under the sampling design $d = (S_r, P_r)$, where S_r is a subset of the universal sample space S and P_r is a probability function defined on S_r, with $\pi_i > 0$ the first order inclusion probability for individual i and π_{ij} the second order probabilities for individuals i and j. Let be

$d_i = 1/\pi_i$ the sampling design weight of unit i. We consider a situation in which U_r and U_v coincide with the population under study U. That is, there are no coverage biases in either the probability or the non-probability sample.

Let us denote with y_i the collected value on the unit i for the target variable y and let $\mathbf{x}_i$ be the observed values for individual i for a vector of covariates $\mathbf{x}$. Both y and $\mathbf{x}$ have been measured in both samples.

The target parameter is the population mean, $\overline{Y} = \frac{1}{N}\sum_U y_i$, that can be estimated from the probability sample using the Horvitz-Thompson estimator:

$$\overline{y}_r = \frac{1}{N}\sum_{i\in s_r} d_i y_i, \tag{1}$$

and from the volunteer sample with the naive estimator:

$$\overline{y}_v = \sum_{i\in s_v} \frac{y_i}{n_v}. \tag{2}$$

If we assume a situation in which there are no non-sampling errors (coverage errors, observation errors, non-response...) the estimator $\overline{y}_r$ is unbiased but if the sample size is small it can lead to estimates with large sampling errors.

Let us consider the variable

$$I_{vi} = \begin{cases} 1 & i \in s_v \\ 0 & i \in U - s_v \end{cases}, \quad i = 1,...,N. \tag{3}$$

The estimator $\overline{y}_v$ is biased (Kim and Wang, 2019) and its bias is given by

$$E_v(\overline{y}_v - \overline{Y}_N) = \frac{1}{f_v} E_v\{Cov(I_v, y)\},$$

where $E_v(.)$ denotes the expectation under the selection mechanism model of the non-probability sample and $f_v = n_v/N$. Thus the mean squared error (MSE) is given by the formula

$$MSE(\overline{y}_v) = \frac{1}{f_v^2} E_v\{Corr(I_v, y)^2\} Var(I_v) Var(y).$$

Therefore, a non-probability sampling where $E_v\{Corr(I_v, y)\} \neq 0$ induces a certain selection bias to the results.

In the next section we will consider how we can estimate the mean population by using a data integration estimator that combine information for these two independent surveys.

3. Methodology in data integration for handling selection bias

3.1. Some previous works

Starting with the work of Elliott and Haviland (2007), these authors consider the problem of combining the two samples by means of a linear combination of the biased and unbiased estimators of the population mean:

$$\overline{y}_{com} = \alpha\overline{y}_r + (1-\alpha)\overline{y}_v.$$

The best estimator, in terms of efficiency, of this combination when the magnitude of the bias is known is given by:

$$\hat{y}_{EH} = \frac{\overline{y}_v \frac{\sigma_r}{n_r} + \overline{y}_r(B^2 + \frac{\sigma_v}{n_v})}{B^2 + \frac{\sigma_r}{n_r} + \frac{\sigma_v}{n_v}}, \tag{4}$$

being $\overline{y}_v$ and $\overline{y}_r$ the sample means, with variances $\frac{\sigma_v}{n_v}$ and $\frac{\sigma_r}{n_r}$ and B the bias of $\overline{y}_v$.
In practice, the bias and variances have to be estimated using the information available from both samples. The bias can be estimated as the difference between the sample means of both samples. In addition, the authors calculate the maximal contribution of the non-probability sample in terms of effective sample size, the role of the non-probability sample size in approaching this limit and the roles of both sample sizes in estimating bias with enough precision. They show that a large probability sample size (1000–10000) is needed for reasonably precise estimates of the remaining bias in initially bias-adjusted non-probability sample estimators.
Other important work is due to Disogra et al. (2011). Their proposal, based on calibration weighting, considers that auxiliary variables needed for calibration weighting must reliably differentiate between the probability sample and the non-probability sample. This calibration method has four steps:

1. Authors do a post-stratification raking calibration of s_r, using a set of demographic and geographical variables.

2. They combine the weighted s_r with the unweighted s_v. The combined sample is then weighted according to the probability sample's benchmarks from the previous step.

3. They compare the answers from early-adopter questions between the probability sample from step 1 to the answers from the blended sample from step 2.

4. They select some minimum number of early adopter questions to include in the raking due in Step 2.

Therefore, this procedure requires a good selection of early adopter questions that are included in the two surveys and that we believe will help to differentiate the samples.
Recently, Kim and Tam (2021) developed two estimators using combined data from probability sampling and non-probability sampling based on the total decomposition:

$$Y = Y_v + Y_c,$$

where $Y_v = \sum_{i\in s_v} y_i = \sum_{i\in U} I_{vi}\, y_i$ and $Y_c = \sum_{i\in U-s_v} y_i = \sum_{i\in U}(1-I_{vi})y_i$. Since y is measured for all units of non-probability sampling, Y_v is known. Therefore, we only have to

estimate Y_c. Authors proposed a first estimator where Y_c is estimated using the expansion estimator based on the probability sample

$$\overline{y}_{DI} = \frac{1}{N}(Y_v + \sum_{i \in s_r} d_i(1 - I_{vi})y_i).$$

In Poisson sampling, the variance of $\overline{y}_{DI}$ is smaller or equal to the variance of $\overline{y}_r$ if a condition on the study variable for simple random sampling without replacement holds. When N is known, Kim and Tam (2021) propose to improve the previous estimator using the following one:

$$\overline{y}_{PDI} = \frac{1}{N}\left(Y_v + (N - n_v)\frac{\sum_{i \in s_r} d_i(1 - I_{vi})y_i}{\sum_{i \in s_r} d_i(1 - I_{vi})}\right).$$

Authors prove that the variance of $\overline{y}_{PDI}$ is smaller than the variance of $\overline{y}_r$ for simple random sampling. They also discuss how to improve the efficiency of this data integration estimator by using ratio and calibration estimation.
Other works in this matter are briefly introduced below.

Fahimi et al. (2015) improve the blended calibration estimator provided by Disogra et al. (2011). Elliot (2009) develop pseudo-weights to create a representative sample using data from the non-probability sample under model assumptions that can be partially tested. With this approach, probability and non-probability samples can be blended, and the resulting sample can be treated as a probability sample with these new pseudo-weights. Dever (2018) proposes a hybrid estimation method based on the combined data file containing probability-based and nonprobability sample cases in a similar way as dual-frame estimation. For this hybrid estimation method , both samples cover the same portion of the population, referred to as common support. The common support assumption is a necessary first step and the authors propose sample matching as the method to evaluate this common support assumption. It is very difficult to make this assumption when we work with web surveys (or social media) and with probabilistic surveys based on population records, as the coverage differences between these samples may be very large and the method cannot be applied. On the other hand, the authors do not solve the problem of the determination of the lambda factor that glues the samples into one data file for population inferences. Wiśniowski et al. (2020) consider a Bayesian approach for integrating a small probability sample with a non-probability sample. They show that considering informative priors based on non-probability data can reduce the variance and mean squared error of the coefficients of a linear model.

Recently, Xi et al. (2022) do an extensive simultation study for comparing various weighting strategies where probability and non-probability samples are combined with weight normalization and raking adjustment. They apply these methods to a teen smoking behaviour survey. Nekrašaitė-Liegė, Čiginas and Krapavickaitė (2022) consider the case of estimating proportions when a non-probabilistic sample and scraped data are available. Some important works (Robbins, Ghosh-Dastidar and Ramchand, 2021;

Rueda et al., 2022) have appeared in which probability and non-probability samples are combined based on the propensity score adjustment technique. In the next section, we explain this technique and how it has been used by these authors.

3.2. Some estimators based on propensity score adjustment

The key concept in a non-probability survey sample is the selection mechanism. This mechanism is usually unknown and requires a suitable prediction model for the inclusion indicator variable. In this context, propensity scores, π_{vi}, can be defined as the probability of the i-th individual of being included in the sample, $P(I_{vi}=1)$, given the characteristics of the unit.

Let $\mathbf{x}$ be a vector of covariates measured in s_v and also in s_r. We make the following assumption:

Assumption 1 (strong ignorability condition): the indicator variable I_v and the study variable y are conditionally independent given x; i.e. $P(I_v=1|\mathbf{x},y)=P(I_v=1|\mathbf{x})$.

We assume that the selection mechanism of s_v verifies Assumption 1 and follows the model:

$$\pi_{vi}=P(I_{vi}=1|\mathbf{x}_i)=p_i(\mathbf{x})=m(\gamma,\mathbf{x}_i),\ \ i=1,...,N, \tag{5}$$

where $m(\cdot)$ is a given function with second continuous derivatives with respect to γ.

We aim to estimate propensity scores using data from pooling both samples. The maximum likelihood estimator of π_{vi} is $m(\hat{\gamma},\mathbf{x}_i)$ where $\hat{\gamma}$ maximizes the pseudo-likelihood (Chen, Li and Wu, 2020):

$$\tilde{l}(\gamma)=\sum_{s_v} log\frac{m(\gamma,\mathbf{x}_i)}{1-m(\gamma,\mathbf{x}_i)}+\sum_{s_r}\frac{1}{\pi_i}log(1-m(\gamma,\mathbf{x}_i)). \tag{6}$$

The estimated propensities $\hat{\pi}_{vi}=m(\hat{\gamma},\mathbf{x}_i)$ are thus used to readjust the propensity bias of the volunteer sample.

Based on these propensities, Robbins et al. (2021) define several estimators integrating the two samples. A first estimator is calculated weighting estimators from each sample:

$$\overline{y}_{RDR1}=\alpha_1\overline{y}_r+(1-\alpha_1)\overline{y}_v, \tag{7}$$

where $\overline{y}_v=\frac{1}{N}\sum_{s_v}y_i/q_i$ with $q_i=\frac{\pi_i*\hat{\pi}_{vi}}{1-\hat{\pi}_{vi}}$ and $\alpha_1=\frac{(\sum_{s_r}\pi_i^{-1})(\sum_{s_v}\hat{\pi}_{vi}^{-2})}{(\sum_{s_r}\pi_i^{-1})(\sum_{s_v}\hat{\pi}_{vi}^{-2})+(\sum_{s_r}\pi_i^{-2})(\sum_{s_v}\hat{\pi}_{vi}^{-1})}$.

For the second estimator, the authors calculate the values $p_i=\pi_i/(1-\hat{\pi}_{vi})$ for all individuals in the joined $s=s_v\cup s_r$ and obtain a simple Horvitz-Thompson type estimator with these new weights:

$$\overline{y}_{RDR2}=\frac{1}{N}\sum_{i\in s}y_i/p_i, \tag{8}$$

Let $\mathbf{x}$ be a set of auxiliary variables, related to y, whose population totals are known. Two calibration estimators are also proposed:

$\bar{y}_{RDR3} = \frac{1}{N}(\sum_{s_v} y_i * w_{1i} + \sum_{s_r} y_i * w_{2i})$ where w_{1i} and w_{2i} are as close as possible to $1/p_i$ fulfilling $T_x = \sum_{\in s_v} w_{1i}\mathbf{x}_i = \sum_{\in s_r} w_{2i}\mathbf{x}_i$ and the estimator:
$\bar{y}_{RDR4} = \alpha_2\bar{y}_r + (1-\alpha_2)\bar{y}_v$ being $\alpha_2 = \frac{(\sum_r w_{i1})(\sum_v w_{i2}^2)}{(\sum_r w_{i1})(\sum_v w_{i2}^2)+(\sum_r w_{i1}^2)(\sum_v w_{i2})}$.

Rueda et al. (2022) propose the combined estimator:

$$\bar{y}_{CPSA} = \alpha_0\bar{y}_r + (1-\alpha_0)\bar{y}_{IPW}, \tag{9}$$

being $\bar{y}_{IPW} = \frac{1}{N}\sum_{s_v} y_i/\hat{\pi}_{vi}$, and $\alpha_0 = \frac{\hat{V}_2}{\hat{V}_1+\hat{V}_2}$ where $\hat{V}_1$ and $\hat{V}_2$ are estimators of the variance of $\bar{y}_r$ and the MSE of $\bar{y}_{IPW}$ respectively. They also propose alternative methods that combine propensity score adjustment and calibration using machine learning predictive algorithms.

Burakauskaitė and Čiginas (2022) consider a few ways on non-probability integration by combining generalized difference estimator and post-stratified calibration estimator with the inverse probability weighted estimating for estimating proportions in the survey on population by religion, native language and ethnicity in Lithuania.

The above methods can reduce bias by using propensity scores to estimate participation rates of non-probability sample units. However, they are sensitive to propensity model misspecifications and can largely increase the variance of the estimators due to extreme weights. A possible way to reduce the effect of extreme weights is the kernel weighting (KW) method Wang et al. (2020) that uses propensity scores as a measure of similarity, and therefore is less sensitive to model misspecification while avoiding the extreme weights that may be produced in propensity score estimation. In the next section we introduce the KW approach to create pseudo-weights for the non-probability sample and propose a new method of integration based on this KW estimator.

4. Estimators based on kernel weighting

The KW method was developed by Wang et al. (2020), and is a method similar to the PSA since both consist of creating pseudo-weights for the non-probability sample using auxiliary variables of a reference probability sample. However, what differentiates them is the way in which these new weights are generated, although as in PSA we will use the estimated propensities to participate in the survey. As it occurred in that case, these propensities can be estimated in different ways, even though the most commonly used one is by means of logistic regression models which may entail several disadvantages for large populations in comparison to modern prediction methods such as machine learning (ML) algorithms. The ML methodology does not require strong parametric model assumptions and therefore is robust to model misspecification. Recently, ML algorithms have been considered in the literature for the treatment of nonprobability samples (see e.g. Ferri-García and Rueda (2020), Buelens et al. (2018), Kern, Li and Wang (2021), Chu and Beaumont (2019), Castro et al. (2021). Their findings showed that ML methods have the potential to remove selection bias in nonprobability samples to a greater extent than logistic regression in some scenarios.

The KW is based on using these propensities to measure the similarity between individuals based on the distributions of the auxiliary variables of the reference sample s_r and the non-probability sample s_v. These similarities will be used as weights for our estimator, after smoothing the distances using kernel functions.

The estimated propensity score for $k \in s_v \cup s_r$ is obtained as

$$\hat{\pi}_k = E_M[\hat{I}_{vk} = 1|\mathbf{x}_k],$$

where M will be one of the mentioned machine learning models to estimate this propensity and

$$\hat{I}_{vk} = \begin{cases} 1 \ \ for \ \ k \in s_v \\ 0 \ \ for \ \ k \in s_r \end{cases}, \quad k \in s_v \cup s_r.$$

Once we have these estimated propensities, we will calculate the distance between the two individuals belonging to the different samples. We define this distance as:

$$d_{ij} = \hat{\pi}_i - \hat{\pi}_j, \ \ i \in s_v, \ \ j \in s_r.$$

This distance between individuals will have a value between -1 and 1. We seek to smooth these values, which is why we use a kernel function centered at zero. There are many alternative kernel functions that can be used (normal function, standard normal, triangular, etc.), see Servy et al. (2006). The closer this distance is to zero, the more similar the individuals are with respect to their auxiliary variables (propensities are estimated depending on the values of the auxiliary variables). Moreover, the more similar the individuals are, the greater the proportion that the KW will assign to the original weight of the reference sample d_{kj} to the i unit of the volunteer sample. This proportion is called the kernel weight, whose expression is as follows:

$$k_{ij} = \frac{K\{d_{ij}/h\}}{\sum_{i \in s_v} K\{d_{ij}/h\}}, \ \ i \in s_v, \ \ j \in s_r,$$

where $K\{\cdot\}$ is a zero-centred kernel function Epanechnikov (1969), and h is the bandwidth corresponding to that kernel function. In addition:

$$\sum_{i \in s_v} k_{ij} = 1, \ \ k_{ij} \in [0,1].$$

The larger the value of the kernel weight k_{ij} is, the more similar the propensities will be among individuals $i \in s_v$ and $j \in s_r$.

Once we have the kernel weights, the pseudo-weights KW can be calculated, w_i^{KW} for $i \in s_v$ which are the sum of the weights of the reference sample d_j, where $j \in s_r$, weighted by the kernel weights k_{ij} for the unit $i \in s_v$:

$$w_i^{KW} = \sum_{j \in s_r} d_j \, k_{ij}, \;\; i \in s_v, \;\; j \in s_r.$$

Therefore a KW estimator for the population mean is:

$$\bar{y}_{KW} = \frac{1}{N} \sum_{i \in s_v} w_i^{KW} y_i,$$

where $\sum_{i \in s_v} w_i^{KW} = \sum_{j \in s_r} d_j$, because of $\sum_{i \in s_v} k_{ij} = 1$.

The KW estimator is consistent if certain regularity conditions are met (see Appendix 1). Furthermore Kern et al. (2021) improve the KW method by pairing it with machine learning, in particular, they considered conditional random forests, model-based recursive partitioning, gradient tree boosting and model-based boosting for estimating the propensities and constructing pseudo-weights. Kernel smoothing is also used by Chen, Yang and Kim (2022) in the case when the study variable of interest is measured only in the non-probability sample. These authors consider mass imputation for the probability sample using the non-probability data as the training set for imputation.

Next, we proceed to present the new proposed method based on KW in two different situations: firstly, if there is no coverage bias for the sample of volunteers, and secondly, when such bias exists.

4.1. Blending the samples with kernel weighting

First, we consider the situation where there is no coverage bias (U_r and U_v are equivalent to the population under study U). In this situation we propose a class of estimators based on both samples:

$$\bar{y}_C = \alpha \bar{y}_r + (1 - \alpha) \bar{y}_{KW}, \tag{10}$$

where α is a nonnegative constant such that $0 \leq \alpha \leq 1$.
We study the asymptotic properties of the proposed estimator under the framework of Isaki and Fuller (1982) in which the properties of estimators are established under a given sequence of populations and a corresponding sequence of random sampling designs.

Theorem 1. *Under assumption given in Appendix 1, the proposed estimator $\bar{y}_C \to Y$ in probability as $N \to \infty$, $n_v \to \infty$, $n_r \to \infty$ with $\frac{n_v}{N} = O(1)$ and $\frac{n_r}{N} = O(1)$.*

Proof. Assumptions 1a and 2a give sufficient conditions for the Horvitz-Thompson estimator $\bar{y}_R$ to be consistent (Isaki and Fuller, 1982). Under these conditions $\bar{y}_R \to \overline{Y}$ in probability as the finite population size $N \to \infty$.

Under assumptions 2a-2c Wang et al. (2020) (Appendix A) proves that $\bar{y}_{KW} \to \overline{Y}$ in probability as the finite population size $N \to \infty$, the survey sample size $n_v \to \infty$ and

the probability sample size $n_r \to \infty$ with $n_c/N = O(1)$. Then it is obtained that $\bar{y}_C \to \alpha\overline{Y} + (1-\alpha)\overline{Y}$ the proposed estimator converges to $\overline{Y}$.
Now, we consider the problem of how select the α parameter. A simple selection for α is to weight each estimator by the weight that sample has in the total sample so that $\alpha_n = n_r/(n_r+n_v)$.

An optimal choice of α can be calculated by minimizing the MSE of $\bar{y}_C$, which is given by

$$MSE(\bar{y}_C) = \alpha^2 V(\bar{y}_r) + (1-\alpha)^2 MSE(\bar{y}_{KW}) + 2\alpha(1-\alpha)E((\bar{y}_r - \overline{Y})(\bar{y}_{KW} - \overline{Y})).$$

As this equation is a quadratic equation of α, its sole extreme is found straightforwardly. The values of α that minimizes this MSE are given by

$$\alpha_{opt} = \frac{MSE(\bar{y}_{KW}) - E((\bar{y}_r - \overline{Y})(\bar{y}_{KW} - \overline{Y}))}{V(\bar{y}_r) + MSE(\bar{y}_{KW}) - 2E((\bar{y}_r - \overline{Y})(\bar{y}_{KW} - \overline{Y}))}. \tag{11}$$

The optimal α_{opt} can be used to define the optimum expression

$$\bar{y}_{Copt} = \alpha_{opt}\bar{y}_r + (1-\alpha_{opt})\bar{y}_{KW}.$$

The optimal coefficient α_{opt} depends on population parameters, which are unknown in practice, and so $\bar{y}_{Copt}$ cannot be calculated.

Though the sampling procedure of the nonprobability and the probability sample can be treated as independent, the estimator $\bar{y}_{KW}$ uses information from both non-probability and probability sample, and therefore can be correlated with $\bar{y}_r$. If we assume that the term $E((\bar{y}_r - \overline{Y})(\bar{y}_{KW} - \overline{Y}))$ is small relative to $MSE(\bar{y}_{KW})$ and $V(\bar{y}_r)$, and denoting by $\hat{V}(\bar{y}_r)$ the Horvitz-Thompson estimator of $V(\bar{y}_r)$ and $\widehat{MSE}(\bar{y}_{KW})$ an estimator for the $MSE(\bar{y}_{KW})$, we can consider the following estimator for the population mean:

$$\bar{y}_{CO} = \frac{\widehat{MSE}(\bar{y}_{KW})}{\widehat{MSE}(\bar{y}_{KW}) + \hat{V}(\bar{y}_r)}\bar{y}_r + \frac{\hat{V}(\bar{y}_r)}{\widehat{MSE}(\bar{y}_{KW}) + \hat{V}(\bar{y}_r)}\bar{y}_{KW}. \tag{12}$$

An estimator for the variance of $\bar{y}_{KW}$ can be obtained by using resampling methods Wolter (2007). By using resampling techniques, one can incorporate aspects of an estimation process into variance calculations that are not easily captured algebraically. Robbins et al. (2021) consider a delete-a-group jackknife for variance estimation when use weighting methods for blending probability and convenience samples. Rafei, Elliott and Flannagan (2022) and Chen et al. (2022) use bootstrap as the method for variance estimation when the study variable of interest is measured only in the non-probability sample. Wang et al. (2020) considered the jackknife method for calculating an estimator of the $V(\bar{y}_{KW})$. The bias of $\bar{y}_{KW}$ can be estimated by $\bar{y}_r - \bar{y}_{KW}$.

4.2. Blending the samples with coverage bias

Web and social media surveys usually have a significant under-coverage bias. Thus, we consider now a more realistic situation where there is also under-coverage bias in the non-probability sample. Chen et al. (2020) highlight the estimation problems in the scenario of having zero propensity scores for certain units in the target population. According to these authors, the severity of the problem depends on the proportion of the uncovered population units and the discrepancies between the two parts of the population in terms of the response variables. Chen (2020) also discusses issues with incomplete sampling frames where units have zero propensity scores and illustrates the danger of applying regular procedures when the sampling frame is incomplete proposing methods to adjust for under coverage bias from the nonprobability sample.

We will consider that U_r covers the entire finite population but the frame U_v be incomplete ($U_v \subset U$). The population of interest, U, may be divided into two mutually exclusive domains, $ab = U_v$ and $a = U \cap U_v^c$. Units in s_r can be divided as $s_r = s_{ra} \cup s_{rab}$, where $s_{ra} = s_r \cap a$ and $s_{rab} = s_r \cap (ab)$.

Following Hartley's idea (Hartley, 1962), we can obtain a combined estimator of $\overline{Y}$ by weighting the estimators obtained from each sample:

$$\overline{y}_H(\eta) = \frac{1}{N}(\hat{Y}_a + \eta \hat{Y}_{ab} + (1-\eta)\hat{Y}_{KW}), \tag{13}$$

where $\hat{Y}_a = \sum_{i \in s_{ra}} d_i y_i$, $\hat{Y}_{ab} = \sum_{i \in s_{rab}} d_i y_i$ and $\hat{Y}_{KW} = \sum_{i \in s_v} w_i^{KW} y_i$ and $0 < \eta < 1$.

Now, we denote as:

$$d_i^\circ = \begin{cases} d_i & \text{if} \quad i \in s_{ra} \\ \eta d_i & \text{if} \quad i \in s_{rab} \\ (1-\eta) w_i^{KW} & \text{if} \quad i \in s_v \end{cases} \tag{14}$$

then

$$\overline{y}_H(\eta) = \frac{1}{N} \sum_{i \in s} d_i^\circ y_i.$$

Theorem 2. *Under the regularity conditions given in Wang et al. (2020) for the sampling design and the propensity scores, the Hartley estimator $\overline{y}_H(\eta)$ is asymptotically unbiased for $\overline{Y}$.*

Proof. Since each domain is estimated by its Horvitz-Thompson estimator, $\hat{Y}_a + \eta \hat{Y}_{ab}$ is an unbiased estimator of $\sum_{i \in a} y_i + \eta \sum_{i \in ab} y_i$, for a given η. Under the regularity conditions given in Wang et al. (2020) the estimator $\hat{Y}_{KW}$ is asymptotically unbiased for $Y_{ab} = \sum_{i \in ab} y_i$, thus the estimator $\overline{y}_H(\eta)$ is asymptotically unbiased for $\overline{Y}$.

Though U_r and U_v are sampled independently, the estimators $\hat{Y}_a + \eta\hat{Y}_{ab}$ and $\hat{Y}_{KW}$ are not independent because, $\hat{Y}_{KW}$ uses information from the probability sample. In the same way as in the previous section, we are going to assume that this dependence is small in relation to the variances of the estimators, and we suppress the covariance term between these two estimators in the calculus of the asymptotic variance of $\bar{y}_H(\eta)$. Under this assumption, the asymptotic variance of the estimator is given by the following expression

$$\begin{aligned} V(\bar{y}_H(\eta)) &= \frac{1}{N^2}(V(\hat{Y}_a + \eta\hat{Y}_{ab}) + V((1-\eta)\hat{Y}_{KW})) \\ &= \frac{1}{N^2}(V(\hat{Y}_a) + \eta^2 V(\hat{Y}_{ab}) + (1-\eta)^2 V(\hat{Y}_{KW})), \end{aligned} \tag{15}$$

where $V(\hat{Y}_a)$ and $V(\hat{Y}_{ab})$ are computed under the sampling design $d = (s_r, p_r)$ and $V(\hat{Y}_{KW})$ under the propensity model π_v.

The choice of the value for η is an important issue. For a fixed value of η, the estimator is simple to implement and gives internal consistency given that the same set of adjusted weights is used for all variables. The value of $\eta = 0.5$ is frequently used in dual frame estimation (Mecatti, 2007). The value of η that minimizes the asymptotic variance in 15 is:

$$\eta_o = \frac{MSE(\hat{Y}_{KW}) - cov(\hat{Y}_a, \hat{Y}_{ab})}{V(\hat{Y}_{ab}) + MSE(\hat{Y}_{KW})}. \tag{16}$$

This value depends on unknown population variances and covariances. By substituting the variances and MSE for its sample based estimators we obtain an estimator that we denote by $\bar{y}_H(opt)$. We note that these modified weights are random variable and their variability needs to be accounted for in standard errors of estimators.

Note. In formula 13, the true population total N is used. It is possible to use an estimator $\hat{N}$ instead of N to construct a type-Hàjek estimator as in the paper of Chen et al. (2020). In our case we would first have to decide which estimator $\hat{N}$ to use. For example based only on the non-probability sample $\hat{N}_1 = \sum_{i \in s_v} w_i^{KW}$, only on the probability sample $\hat{N}_2 = \sum_{i\ ins_r} d_i$ or some estimator based on the two samples. This choice can influence the biasness and the efficiency of the proposed estimator, and adds one more difficulty to the problem.

5. Simulation studies

We have conducted a simulation study to compare the efficiency of some of the proposed estimators based on KW. We are interested in comparing those estimators with some alternative estimators defined in Section 3, in the effect of the machine learning algorithm used in KW, in the effect of the kernel function used in the construction of

KW pseudo-weights and also in the effect of considering coverage bias. In order to illustrate that the superiority of some estimators compared to others depends on the data, we define different setups based on different artificial populations and different sampling strategies.

5.1. Populations and setups

We consider a finite population of size N = 500000. The variables of interest were designed with the objective of having various types of relationships with the covariates and the propensities. We consider 8 auxiliary variables x, 2 variable of interest y and a variable π_{vi} which indicates the probability of being included in the non-probability sample. All of them were simulated as follows:

1. The covariates x_1, x_3, x_5 and x_7 followed a Bernoulli distribution with $p = 0.5$, and x_2, x_4, x_6 and x_8 followed normal distributions with standard deviation of one and a mean parameter of 0 or 2, depending on the value of the previous Bernoulli variable. That is to say, in order to calculate x_2 we relied on the variable x_1 and if this variable was equal to 1, then the mean would be 2, or if the variable was equal to 0, then the mean would be 0. The same procedure was followed for the rest of the variables. The propensity models were fitted using all of the 8 auxiliary variables.

2. The non-probability samples were drawn with a Poisson sampling design where the inclusion probability depends on variables x_5, x_6, x_7 y x_8 as:

$$ln\left(\frac{\pi_{vi}}{1-\pi_{vi}}\right) = -0.5 + 2.5(x_{5i} = 1) + \sqrt{2 \cdot 3.141593}\, x_{6i} x_{8i} - 2.5(x_{7i} = 1), \quad i \in U. \tag{17}$$

3. The target variables were created in order to have different relationships with the covariates and the propensities were simulated according to the formulas:

$$\begin{aligned} y_{1i} &= N(8,2) + 3(x_{5i} = 1) + 5\pi_i, \quad i \in U; \\ y_{2i} &= \begin{cases} 1 & \text{if } y_{1i} > 14.46 \\ 0 & \text{if } y_{1i} \leq 14.46 \end{cases}, \quad i \in U. \end{aligned} \tag{18}$$

The threshold of 14.46 was chosen because it is equivalent to the median of the variable y_1.

We considered three setups. In the first setup the probability sample was drawn by simple random sampling without replacement (SRSWOR) from the full population; in the second setup the probability sample was drawn with stratified random sampling by the auxiliar variable x_7 and considering an allocation by strata of 1/3 and 2/3; in the third setup, the probability sample was selected with Midzuno sampling where the probabilities were proportional to a variable following a normal

distribution with a mean parameter dependent on the value of the auxiliar variable x_7 and a standard deviation of 0.5.

The aim of the described selection mechanism was to create weights with large variability. As a result, the mean propensity is 0.7050, with a standard deviation of 0.3792, and thus a coefficient of variation of 0.5379. The histogram of propensities $\pi_{vi}, i \in U$, is provided in figure 1.

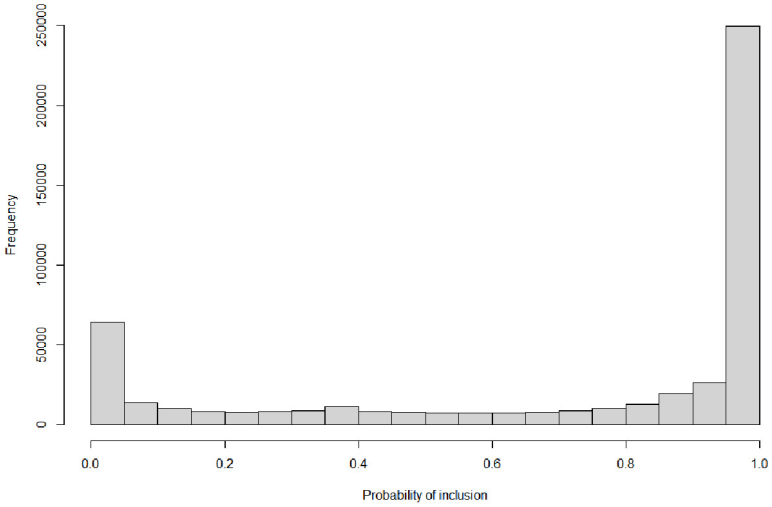


Figure 1. *Histogram of the population propensities.*

5.2. The simulation procedure

The first simulation study evaluates the performance of some estimators for $\overline{Y}$ when there is selection bias in the estimates. We focused on the proposed estimator discussed in the paper, $\overline{y}_{CO}$, and we compared it with others estimators based on propensities. As a reference estimator we have considered the naive estimator that weights the estimators simply by their sizes $\overline{y}_{REF} = \frac{n_r}{N_r}\overline{y}_r + \frac{n_v}{N_v}\overline{y}_v$. We also evaluate the estimators $\overline{y}_{RDR1}$ (7), $\overline{y}_{RDR2}$ (8) and $\overline{y}_{CPSA}$ (9) that do not use calibration.

We considered the XGBoost (Chen and Guestrin, 2016) algorithm among several machine learning approaches for estimating the propensities in all estimators. This algorithm builds decision trees ensembles that optimize an objective function via gradient tree boosting (Friedman, 2001). Literature shows that PSA with gradient boosting machines provides better results than other machine learning approaches (Lee, Lessler and Stuart, 2010, 2011; McCaffrey, Ridgeway and Morral, 2004, 2013; Ferri-García and Rueda, 2020; Rueda et al., 2022). The method depends on several hyperparameters for a proper functioning and in order to avoid overfitting. We have considered the following hyperparameters: the number of trees forming the ensemble (50, 100 or 150), the weight

shrinkage applied after each boosting step (0.3 or 0.4), the maximum number of splits that each tree can contain (1, 2 or 3), the proportion of variables used in each step (0.6 or 0.8) and the proportion of data used in each step (0.5, 0.75 or 1).

For each setup we select 500 probability samples of size $n_r = 250$ and 500 non-probability samples of sizes $n_v = 500; 1000; 2000$. We compute the Monte Carlo relative bias of the estimators:

$$|RB| = \frac{1}{B}\sum_{b=1}^{B}\frac{|\overline{y}_b - \overline{Y}|}{\overline{Y}} \cdot 100, \tag{19}$$

and the Monte Carlo root mean square relative error (RMSRE):

$$RMSRE = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\frac{\overline{y}_b - \overline{Y}}{\overline{Y}}\right)^2} \cdot 100. \tag{20}$$

where B is the number of iterations, and $\overline{y}_b$ is an estimate of $\overline{Y}$, by the method under study, computed for the b-th sample.

We also examine the behaviour of variance estimators. We consider the jackknife method used in Wang et al. (2020) to account for all sources of variability. The performance of a variance estimator along with the point estimator $\overline{y}_i$ is assessed by the length of the intervals obtained at 95% confidence level and their real coverage.

Variance estimators for $\overline{y}_{KW}$ is also calculated based on bootstrap methods. We have obtained similar results for RB and RMSRE for the proposed estimator $\overline{y}_{CO}$ and we observed that the behaviour with respect to the other estimators is barely influenced by the variance estimation method used. In the work only the results of the jackkniffe method are shown.

The simulation study has been carried out using the statistical software R, and for its implementation we have needed the use of specific packages of the area, such as NonProbEst (Castro, Ferri and Rueda, 2020), KWML (Kern et al., 2021), sampling (Tillé and Matei, 2021) and caret (Kuhn et al., 2022).

5.3. Results

Tables 1 and 2 contain the simulation results for y_1 and y_2 respectively for the three setups considering different sample sizes. In all setups, as expected, the proposed estimator with gradient boosting and kernel weighting ($\overline{y}_{CO}$) provides lower values of both |RB| and RMSRE. The second best estimator is $\overline{y}_{CPSA}$, which obtains results similar to the first and with the rest of the estimators we obtain higher values of the |RB| and RMSRE. It is also observed that the behaviour pattern in terms of reduction |RB| and RMSRE is similar in the three sample designs considered for the probabilistic sample.

Table 1. *Monte Carlo bias and root mean square relative error. Variable* y_1.

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
Simple random sampling without replacement						
$\bar{y}_{REF}$	4.772	4.847	4.732	4.795	4.801	4.872
$\bar{y}_{RDR1}$	3.081	3.231	2.736	2.895	2.770	2.952
$\bar{y}_{RDR2}$	3.246	3.389	2.888	3.045	2.907	3.088
$\bar{y}_{CPSA}$	1.251	1.554	1.197	1.512	1.341	1.663
$\bar{y}_{CO}$	1.173	1.457	1.232	1.559	1.213	1.576
Stratified sampling						
$\bar{y}_{REF}$	4.800	4.880	4.864	4.942	4.730	4.788
$\bar{y}_{RDR1}$	2.998	3.190	2.913	3.114	2.752	2.910
$\bar{y}_{RDR2}$	3.651	3.788	3.601	3.746	3.435	3.546
$\bar{y}_{CPSA}$	1.448	1.798	1.595	2.003	1.326	1.665
$\bar{y}_{CO}$	1.224	1.521	1.322	1.671	1.162	1.431
Midzuno sampling						
$\bar{y}_{REF}$	4.771	4.845	4.766	4.827	4.735	4.792
$\bar{y}_{RDR1}$	3.100	3.257	2.801	2.947	2.766	2.912
$\bar{y}_{RDR2}$	3.381	3.520	3.122	3.250	3.069	3.198
$\bar{y}_{CPSA}$	1.219	1.526	1.261	1.554	1.2390	1.573
$\bar{y}_{CO}$	1.010	1.393	1.141	1.412	1.124	1.425

Table 2. *Monte Carlo bias and root mean square relative error. Variable* y_2.

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
Simple random sampling without replacement						
$\bar{y}_{REF}$	16.229	16.653	16.199	16.547	16.396	16.775
$\bar{y}_{RDR1}$	10.025	10.769	9.130	9.809	9.188	9.971
$\bar{y}_{RDR2}$	10.650	11.367	9.600	10.281	9.605	10.398
$\bar{y}_{CPSA}$	5.188	6.406	5.136	6.400	5.759	7.176
$\bar{y}_{CO}$	4.538	5.665	4.681	5.920	5.343	6.642
Stratified sampling						
$\bar{y}_{REF}$	16.317	16.738	16.703	17.133	16.222	16.537
$\bar{y}_{RDR1}$	11.010	11.734	11.012	11.772	10.518	11.068
$\bar{y}_{RDR2}$	12.819	13.412	12.821	13.447	12.326	12.785
$\bar{y}_{CPSA}$	5.647	7.110	5.866	7.613	5.126	6.408
$\bar{y}_{CO}$	5.119	6.444	5.198	6.704	4.612	5.684
Midzuno sampling						
$\bar{y}_{REF}$	16.421	16.829	16.248	16.581	16.690	17.030
$\bar{y}_{RDR1}$	10.738	11.437	9.866	10.512	10.182	10.824
$\bar{y}_{RDR2}$	11.634	12.271	10.771	11.382	11.020	11.632
$\bar{y}_{CPSA}$	5.246	6.652	5.052	6.206	5.632	7.004
$\bar{y}_{CO}$	4.706	5.903	4.490	5.583	4.965	6.163

Table 3. *Confidence invervals' real coverage and length. Variable* y_1.

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
Simple random sampling without replacement						
$\bar{y}_{REF}$	0.000	0.481	0.000	0.453	0.000	0.438
$\bar{y}_{RDR1}$	0.108	0.517	0.168	0.492	0.164	0.475
$\bar{y}_{RDR2}$	0.088	0.522	0.146	0.505	0.146	0.490
$\bar{y}_{CPSA}$	0.956	0.853	0.962	0.854	0.918	0.855
$\bar{y}_{CO}$	0.962	0.811	0.960	0.807	0.928	0.781
Stratified sampling						
$\bar{y}_{REF}$	0.002	0.503	0.002	0.477	0.000	0.463
$\bar{y}_{RDR1}$	0.190	0.552	0.178	0.525	0.178	0.509
$\bar{y}_{RDR2}$	0.054	0.533	0.052	0.507	0.030	0.494
$\bar{y}_{CPSA}$	0.944	0.939	0.898	0.948	0.954	0.948
$\bar{y}_{CO}$	0.958	0.844	0.906	0.822	0.952	0.788
Midzuno sampling						
$\bar{y}_{REF}$	0.000	0.488	0.000	0.462	0.000	0.445
$\bar{y}_{RDR1}$	0.124	0.529	0.146	0.505	0.140	0.487
$\bar{y}_{RDR2}$	0.090	0.528	0.090	0.509	0.094	0.493
$\bar{y}_{CPSA}$	0.958	0.886	0.962	0.887	0.956	0.886
$\bar{y}_{CO}$	0.952	0.820	0.964	0.808	0.950	0.769

Table 4. *Confidence invervals' real coverage and length. Variable* y_2.

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
Simple random sampling without replacement						
$\bar{y}_{REF}$	0.008	0.076	0.006	0.070	0.002	0.066
$\bar{y}_{RDR1}$	0.276	0.077	0.286	0.070	0.236	0.066
$\bar{y}_{RDR2}$	0.242	0.077	0.252	0.071	0.232	0.068
$\bar{y}_{CPSA}$	0.968	0.130	0.954	0.130	0.930	0.131
$\bar{y}_{CO}$	0.950	0.118	0.954	0.119	0.904	0.116
Stratified sampling						
$\bar{y}_{REF}$	0.018	0.080	0.002	0.074	0.002	0.071
$\bar{y}_{RDR1}$	0.198	0.079	0.174	0.072	0.132	0.069
$\bar{y}_{RDR2}$	0.108	0.078	0.084	0.071	0.052	0.068
$\bar{y}_{CPSA}$	0.944	0.139	0.924	0.140	0.976	0.140
$\bar{y}_{CO}$	0.932	0.126	0.916	0.121	0.944	0.114
Midzuno sampling						
$\bar{y}_{REF}$	0.010	0.077	0.002	0.071	0.002	0.068
$\bar{y}_{RDR1}$	0.232	0.077	0.232	0.071	0.162	0.067
$\bar{y}_{RDR2}$	0.168	0.078	0.178	0.072	0.126	0.068
$\bar{y}_{CPSA}$	0.950	0.133	0.988	0.134	0.958	0.134
$\bar{y}_{CO}$	0.950	0.122	0.960	0.118	0.924	0.115

Table 5. *Monte Carlo bias and root mean square relative error of estimators changing the ML method. Variable y_1.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
Simple random sampling without replacement						
$\overline{y}_{CO}$	1.156	1.406	1.162	1.428	1.260	1.570
$\overline{Y}_{CO-NNET}$	1.165	1.422	1.243	1.521	1.317	1.676
$\overline{Y}_{CO-K}$	1.165	1.418	1.165	1.438	1.270	1.610
$\overline{Y}_{CO-LR}$	1.197	1.468	1.279	1.568	1.339	1.695
Stratified sampling						
$\overline{y}_{CO}$	1.250	1.547	1.261	1.595	1.257	1.578
$\overline{Y}_{CO-NNET}$	1.389	1.713	1.379	1.773	1.474	1.829
$\overline{Y}_{CO-K}$	1.234	1.527	1.250	1.582	1.240	1.550
$\overline{Y}_{CO-LR}$	1.467	1.814	1.477	1.891	1.557	1.923
Midzuno sampling						
$\overline{y}_{CO}$	1.254	1.567	1.191	1.478	1.307	1.615
$\overline{Y}_{CO-NNET}$	1.331	1.665	1.277	1.605	1.490	1.884
$\overline{Y}_{CO-K}$	1.272	1.592	1.203	1.495	1.337	1.658
$\overline{Y}_{CO-LR}$	1.382	1.732	1.313	1.650	1.529	1.929

Table 6. *Monte Carlo bias and root mean square relative error of estimators changing the ML method. Variable y_2.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
Simple random sampling without replacement						
$\overline{y}_{CO}$	5.144	6.264	4.932	6.105	5.115	6.378
$\overline{Y}_{CO-NNET}$	5.510	6.760	5.315	6.554	5.618	6.840
$\overline{Y}_{CO-K}$	5.107	6.278	5.023	6.240	5.057	6.334
$\overline{Y}_{CO-LR}$	5.739	7.011	5.558	6.903	5.842	7.101
Stratified sampling						
$\overline{y}_{CO}$	5.045	6.334	5.151	6.566	5.200	6.455
$\overline{Y}_{CO-NNET}$	5.449	6.848	5.870	7.472	6.058	7.461
$\overline{Y}_{CO-K}$	4.967	6.312	5.240	6.716	5.553	6.837
$\overline{Y}_{CO-LR}$	5.593	7.028	5.939	7.508	6.197	7.615
Midzuno sampling						
$\overline{y}_{CO}$	4.781	5.868	4.9700	6.309	5.137	6.327
$\overline{Y}_{CO-NNET}$	5.290	6.467	5.683	6.972	5.432	6.698
$\overline{Y}_{CO-K}$	4.922	6.078	5.175	6.454	5.086	6.292
$\overline{Y}_{CO-LR}$	5.520	6.717	5.807	7.124	5.592	6.916

Tables 3 and 4 show the real coverages and lengths of the corresponding 95% confidence intervals. The coverage of intervals based on estimators $\overline{y}_{REF}$, $\overline{y}_{RDR1}$ and $\overline{y}_{RDR2}$ are very low, as expected, due to the bias in the estimates. On the contrary, the proposed es-

Table 7. *Confidence invervals' real coverage and length changing the ML method. Variable* y_1.

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
Simple random sampling without replacement						
$\overline{y}_{CO}$	0.974	0.812	0.948	0.805	0.924	0.780
$\overline{Y}_{CO-NNET}$	0.970	0.830	0.956	0.839	0.918	0.858
$\overline{Y}_{CO-K}$	0.970	0.823	0.960	0.820	0.936	0.821
$\overline{Y}_{CO-LR}$	0.970	0.854	0.956	0.867	0.916	0.876
Stratified sampling						
$\overline{y}_{CO}$	0.946	0.845	0.930	0.820	0.916	0.784
$\overline{Y}_{CO-NNET}$	0.926	0.905	0.932	0.915	0.926	0.929
$\overline{Y}_{CO-K}$	0.954	0.854	0.938	0.849	0.936	0.849
$\overline{Y}_{CO-LR}$	0.918	0.936	0.924	0.951	0.920	0.963
Midzuno sampling						
$\overline{y}_{CO}$	0.914	0.823	0.952	0.804	0.918	0.770
$\overline{Y}_{CO-NNET}$	0.922	0.860	0.940	0.875	0.892	0.882
$\overline{Y}_{CO-K}$	0.930	0.835	0.952	0.827	0.912	0.829
$\overline{Y}_{CO-LR}$	0.918	0.893	0.950	0.901	0.908	0.911

timator $\overline{y}_{CO}$ and $\overline{y}_{CPSA}$ have good performance, having the intervals a real coverage close to the nominal coverage. With respect to the length of the intervals, as we expected, the $\overline{y}_{CO}$ estimator is the one with the shortest length for all types of sampling considered, sample sizes and type of variable. The KW is intended to reduce variance and indeed it succeeds for these scenarios and variables.

Table 8. *Confidence intervals' real coverage and length changing the ML method. Variable* y_2.

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
Simple random sampling without replacement						
$\overline{y}_{CO}$	0.940	0.118	0.938	0.120	0.926	0.119
$\overline{Y}_{CO-NNET}$	0.906	0.124	0.926	0.125	0.900	0.126
$\overline{Y}_{CO-K}$	0.954	0.120	0.944	0.120	0.938	0.120
$\overline{Y}_{CO-LR}$	0.920	0.128	0.916	0.130	0.900	0.130
Stratified sampling						
$\overline{y}_{CO}$	0.950	0.127	0.928	0.120	0.898	0.116
$\overline{Y}_{CO-NNET}$	0.944	0.138	0.930	0.139	0.942	0.139
$\overline{Y}_{CO-K}$	0.960	0.128	0.952	0.127	0.950	0.127
$\overline{Y}_{CO-LR}$	0.938	0.140	0.920	0.140	0.946	0.143
Midzuno sampling						
$\overline{y}_{CO}$	0.952	0.120	0.934	0.120	0.900	0.114
$\overline{Y}_{CO-NNET}$	0.964	0.130	0.914	0.130	0.958	0.133
$\overline{Y}_{CO-K}$	0.962	0.123	0.944	0.121	0.958	0.122
$\overline{Y}_{CO-LR}$	0.958	0.134	0.930	0.135	0.952	0.137

5.4. Influence of the machine learning method

In the previous simulation we used gradient boosting machine as a machine learning method, but different methods can be used. In this case we are going to make a comparison of the most used machine learning methods to see if the results are influenced by them. Specifically, we are going to compare neural networks (NNET), K-nearest neighbours (K) and logistic regression (LR) with respect to gradient boosting machine for qualitative and quantitative variables y_1 and y_2 considering the three types of sampling and for the different sample sizes. The results obtained in the comparative study can be seen in the Tables 5, 6, 7 and 8.

When comparing the |RB| and the RMSRE values for y_1 for all sample sizes (Table 5), we can see that in simple random sampling and Midzuno sampling the smallest values are found for $\overline{y}_{CO}$, in the case of stratified sampling, the smallest values are found in $\overline{Y}_{CO-K}$. For y_2 (Table 6) the results obtained for the gradient boosting machine and K-nearest neighbours method are similar if we compare the |RB| and the RMSRE values. When looking at the Tables 7 and 8 for y_1 it can be observed that the greatest coverage (0.91-0.97) obtained is given in the case of the gradient boosting machine and K-nearest neighbours methods. For y_2 the K-nearest neighbours method obtains the greatest coverage (0.93-0.96). With respect to the length of the confidence interval, gradient boosting machine obtains the smallest values and logistic regression model obtains the largest. The performance of the logistic regression was to be expected since the propensities do not depend on all the covariates and there is an error in the propensity model specification.

Table 9. *Monte Carlo bias and root mean square relative error of estimators changing the kernel. Variable* y_1.

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
	Simple random sampling without replacement					
$\overline{y}_{CO}$	1.160	1.448	1.144	1.437	1.261	1.578
$\hat{\overline{Y}}_{CO-SN}$	1.164	1.449	1.140	1.435	1.264	1.577
$\hat{\overline{Y}}_{CO-TSN}$	1.161	1.451	1.145	1.437	1.261	1.577
	Stratified sampling					
$\overline{y}_{CO}$	1.245	1.573	1.414	1.734	1.210	1.492
$\hat{\overline{Y}}_{CO-SN}$	1.250	1.579	1.389	1.703	1.206	1.492
$\hat{\overline{Y}}_{CO-TSN}$	1.256	1.597	1.389	1.719	1.110	1.489
	Midzuno sampling					
$\overline{y}_{CO}$	1.221	1.540	1.229	1.513	1.312	1.631
$\hat{\overline{Y}}_{CO-SN}$	1.220	1.536	1.232	1.518	1.308	1.626
$\hat{\overline{Y}}_{CO-TSN}$	1.230	1.548	1.231	1.518	1.320	1.632

Table 10. *Monte Carlo bias and root mean square relative error of estimators changing the kernel. Variable y_2.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
Simple random sampling without replacement						
$\bar{y}_{CO}$	4.641	5.762	4.839	6.070	5.012	6.378
$\hat{\bar{Y}}_{CO-SN}$	4.567	5.679	4.832	6.088	5.002	6.335
$\hat{\bar{Y}}_{CO-TSN}$	4.627	5.776	4.783	6.041	5.040	6.409
Stratified sampling						
$\bar{y}_{CO}$	5.215	6.627	4.902	6.175	5.069	6.298
$\hat{\bar{Y}}_{CO-SN}$	5.199	6.612	4.991	6.250	5.064	6.332
$\hat{\bar{Y}}_{CO-TSN}$	5.271	6.631	4.988	6.230	5.099	6.377
Midzuno sampling						
$\bar{y}_{CO}$	4.657	5.873	5.122	6.274	4.966	6.211
$\hat{\bar{Y}}_{CO-SN}$	4.736	5.896	5.202	6.311	5.014	6.263
$\hat{\bar{Y}}_{CO-TSN}$	4.617	5.870	5.220	6.375	4.993	6.256

Table 11. *Confidence intervals' real coverage and length changing the kernel. Variable y_1.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
Simple random sampling without replacement						
$\bar{y}_{CO}$	0.946	0.812	0.962	0.807	0.918	0.782
$\hat{\bar{Y}}_{CO-SN}$	0.956	0.814	0.966	0.811	0.920	0.790
$\hat{\bar{Y}}_{CO-TSN}$	0.950	0.812	0.968	0.810	0.918	0.789
Stratified sampling						
$\bar{y}_{CO}$	0.946	0.843	0.912	0.828	0.948	0.785
$\hat{\bar{Y}}_{CO-SN}$	0.954	0.851	0.930	0.831	0.936	0.791
$\hat{\bar{Y}}_{CO-TSN}$	0.932	0.843	0.932	0.831	0.946	0.793
Midzuno sampling						
$\bar{y}_{CO}$	0.930	0.821	0.958	0.807	0.912	0.777
$\hat{\bar{Y}}_{CO-SN}$	0.942	0.825	0.960	0.813	0.910	0.782
$\hat{\bar{Y}}_{CO-TSN}$	0.932	0.821	0.958	0.810	0.914	0.785

5.5. Influence of the kernel function

In the previous simulations we used the tringular distribution as kernel function in the construction of KW pseudo-weights, but different distributions can be used. In this case we are going to make a comparison of the distribution implemented in the R package Boosted Kernel Weighting (Kern et al., 2021) to see if the results are influenced by them. Specifically, we are going to compare triangular, standard normal (SN) and truncated standard normal (TSN) for qualitative and quantitative variables y_1 and y_2 considering

Table 12. *Confidence intervals' real coverage and length changing the kernel. Variable y_2.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
Simple random sampling without replacement						
$\bar{y}_{CO}$	0.944	0.118	0.940	0.118	0.932	0.117
$\hat{\bar{Y}}_{CO-SN}$	0.958	0.120	0.934	0.121	0.930	0.119
$\hat{\bar{Y}}_{CO-TSN}$	0.956	0.119	0.936	0.121	0.932	0.119
Stratified sampling						
$\bar{y}_{CO}$	0.924	0.125	0.954	0.121	0.922	0.114
$\hat{\bar{Y}}_{CO-SN}$	0.926	0.126	0.944	0.122	0.926	0.117
$\hat{\bar{Y}}_{CO-TSN}$	0.938	0.126	0.944	0.122	0.916	0.117
Midzuno sampling						
$\bar{y}_{CO}$	0.952	0.121	0.942	0.119	0.920	0.117
$\hat{\bar{Y}}_{CO-SN}$	0.950	0.123	0.942	0.122	0.918	0.117
$\hat{\bar{Y}}_{CO-TSN}$	0.948	0.122	0.936	0.120	0.916	0.117

the three types of sampling and for the different sample sizes. The results obtained in the comparative study can be seen in the Tables 9, 10, 11 and 12.

Table 13. *Monte Carlo bias and root mean square relative error of estimators with coverage bias. Variable y_1.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
$\bar{y}_{REF}$	5.541	5.615	5.581	5.649	5.554	5.619
$\bar{y}_{RDR1}$	3.279	3.427	3.295	3.421	3.175	3.304
$\bar{y}_{RDR2}$	3.233	3.409	3.198	3.358	2.999	3.166
$\bar{y}_{CPSA}$	1.267	1.574	1.213	1.535	1.220	1.543
$\bar{y}_{CO}$	1.258	1.563	1.209	1.529	1.204	1.520
$\bar{y}_H(opt)$	1.195	1.486	1.125	1.426	1.132	1.446

Table 14. *Monte Carlo bias and root mean square relative error of estimators with coverage bias. Variable y_2.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
$\bar{y}_{REF}$	19.689	20.085	19.689	20.085	19.689	20.085
$\bar{y}_{RDR1}$	12.210	12.828	12.210	12.828	12.210	12.828
$\bar{y}_{RDR2}$	12.118	12.794	12.118	12.794	12.118	12.794
$\bar{y}_{CPSA}$	5.359	6.623	5.359	6.623	5.359	6.623
$\bar{y}_{CO}$	5.375	6.648	5.375	6.648	5.375	6.648
$\bar{y}_{H(opt)}$	5.258	6.453	5.258	6.453	5.258	6.453

The values of |RB| and the RMSRE are similar for the kernel functions used, so we can say that there is no influence of the kernel function in this study. Regarding coverage, we see that in all cases it is quite good, moving around 0.91–0.96, obtaining the shortest length of the interval in most cases in the $\bar{y}_{CO}$ estimator.

5.6. Results under coverage bias

In order to check the behaviour of the Hartley estimator $\bar{y}_H(opt)$, proposed in section 4.2, we have repeated the previous simulation but now we include a mechanism to reproduce coverage bias in our simulation. This context is compared with the same estimators considered in the first simulation.

The probability sample is selected by SRSWOR from the full population but the non-probability sample is now selected from a frame U_v created from the population U containing only individuals whose variable $x_5 = 1$ (related to target variables).

In Tables 13 and 14 values of |RB| and the RMSRE can be seen for each of the considered estimators.

As expected, all the estimators considered now have greater bias than in the previous simulation. We observe that the estimators $\bar{y}_{CPSA}$ and $\bar{y}_{CO}$ continue to be better than the other PSA-based estimators in terms of |RB| and RMSRE reduction. As expected, the estimator based on dual frames, $\bar{y}_H(opt)$, is the one that produces estimates with less |RB|, and consequently is also able to reduce the RMSRE compared to its competitors.

6. Discussion

In the last decade, survey research has witnessed the surge of non-probability sampling as a feasible alternative to probability sampling. In theory, the superiority of probability sampling should be clear, as it has a theoretical basis in design-based inference allowing for unbiased estimation of population parameters along with the calculation of exact sampling error. However, they are very expensive and usually have small sizes. Non-probability samples can offer some advantages in that sense, as they can be deployed in many relatively inexpensive ways, but they lack an underlying mathematical theory given their usual lack of design. This is troublesome with respect to achieving accuracy and representativeness for estimates derived from such samples.

Given their potential, many efforts have been undertaken in recent years to combine both probability and nonprobability samples to produce a single inference which may be able to overcome the limitations of each method, resulting in a rich literature on data integration in finite populations. Most of this literature is based on considering a framework where the variables of interest have not been observed in the probability sample. In this paper, we have considered the problem of observed study variables in both the non-probability sample and the probability sample, in presence of auxiliary information.

Since both samples contain the same variables, we propose a methodology to combine two surveys based on probability and non-probability samples with the help of ma-

chine learning algorithms, in order to obtain reliable estimations with small variance. We have introduced a general class of estimators, based on the kernel weighting method, and studied theoretically their bias properties. Using simulations we have also compared the proposed estimators with other methods for integrating probability and non-probability samples developed in the literature in different simulation setups, both in terms of |RB| and RMSRE.

The simulation study indicates that |RB| and RMSRE of estimators can be reduced when combining the probability and the non-probability sample using the KW method proposed here in the case where there is a relationship between the variable of interest and the participation probability. We also observed that the choice of the ML method used for propensity predictions is very important and can influence the estimates obtained. However, the kernel function in the construction of KW pseudo-weights does not influence the estimates obtained. From our simulation study we also deduce that in case the sample of volunteers has a coverage bias, it is appropriate to use an estimator based on dual frames that allows this bias to be treated as well.

These methods can be implemented using freely available statistical packages such as R. The R code used for the simulation study and the computation of the results are available on request. However, the computational cost of resampling should be mentioned. Many of the proposed methods rely on variance estimation techniques which involve resampling. For each iteration, a new model has to be trained and the calculations have to be repeated, considerably slowing down the process. Therefore, they should be avoided when execution time is of the essence and many variables are involved.

Some other papers (Elliot (2009), Dever (2018)) also combine the pseudo-weighted nonprobability and probability samples first and estimate the finite population mean from the combined sample. When pseudo-weighted samples are combined, the assigned weights only depend on the sample sizes, the design weights and the estimated propensities, which do not depend on the variable under study. Thus, the same weights are used to make estimates for all variables, but for some variables the procedure may not be able to eliminate voluntariness biases. On the contrary, the method that we propose depends on each variable under study, and takes into account the voluntariness bias that may be important for those variables that are correlated with the probability of participating in the survey of volunteers, which is the case that interests us.

In our proposal we have considered non parametric methods to estimate the underlying propensity model that reflect the self-selection process, which provides added flexibility over logistic regression-based methods. Some recent works also use non-parametric methods to make inferences for non-probability samples. Chen et al. (2022) use kernel smoothing while Yang, Kim and Hwang (2021) use nearest neighbor for mass imputation for the probability sample using the non-probability data as the training set. Our method differs from these works fundamentally in two aspects: in our case the variable under study is observed in the two samples, and we use the inverse propensity weighting methodology while they use mass imputation.

Our advice to practitioners is that the use of probability samples remains essential to obtain reliable estimates based on an accepted theory such as sampling theory (Beaumont, 2020), but complementing the probability sample with a non-probability sample can serve a means to reduce the errors in the estimates.

There is a lot of room for future research to improve estimation by mean integration: other similarity measures and other weighting adjustment methods such as weight smoothing for multipurpose surveys (Ferri-García et al., 2022) can be considered. In this work only the estimation of means and totals has been considered, but the method can be applied, with certain adjustments, to the case of other non-linear parameters such as distribution functions or quantiles. In addition, new alternative methods for estimation from a nonprobability sample continue to emerge. Liu and Valliant (2023) introduces one method of weighting that assign a unit in the nonprobability sample the weight from its matched case in the probability sample. These new methods can be used as an alternative to kernel weighting to build estimators similar to our proposal. These issues will be future research topics.

Acknowledgments

This work is part of grant PID2019-106861RB-I00 funded by MCIN/ AEI/10.13039/501100011033, by grant PDC2022-133293-I00 funded by MCIN/AEI/10.13039/5011000 11033 and the European Union "NextGenerationEU"/PRTR"1 and the grant FEDER C-EXP-153-UGR23 funded by Consejería de Universidad, Investigación e Innovación and by ERDF Andalusia Program 2021-2027. The research was also partially supported from IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033.

Conflict of interest

The authors declare no potential conflict of interests.

References

Bailey, M.A. (2022). Comments on "Statistical inference with non-probability survey samples" - Non-probability samples: An assessment and way forward. Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00003-eng.htm.

Beaumont, J.F. (2020). Are probability surveys bound to disappear for the production of official statistics?. *Survey Methodology, Statistics Canada*, 46, 1. http://www.statcan.gc.ca/pub/12-001-x/2020001/article/00001-eng.htm.

Buelens, B., Burger, J. and vanden Brakel, J.A. (2018). Comparing inference methods for non-probability samples. *International Statistical Review* 86(2), 322-343.

Burakauskaitė, I., Čiginas, A. (2022). Non-probability sample integration in the survey of lithuanian census. *Workshop on Survey Statistics, Tartu, 2022*. http://isi-iass.org/home/wp-content/uploads/I.Burakauskaite_A.Ciginas_presentation.pdf.

Castro, L., Ferri, R. and Rueda, M.M. (2020). *N*onProbEst: Estimation in Nonprobability Sampling. https://CRAN.R-project.org/package=NonProbEst.

Castro, L., Rueda, M., Ferri-García, R. and Hernando-Tamayo, C. (2021). On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures. *M*athematics 9, 2991. https://doi.org/10.3390/math9232991

Castro L., Rueda M. and Ferri-García R. (2022). Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys *J*ournal of Computational and Applied Mathematics, 404, 113414.

Chen, S., Yang, S. and Kim, J.K. (2022). Nonparametric Mass Imputation for Data Integration. *Journal of Survey Statistics and Methodology*, 10(1), 1-24.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA* 785-794.

Chen, Y. (2020). Statistical Analysis with Non-Probability Survey Samples. Ph Thesis

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021.

Chu, K.C.K. and Beaumont, J.F. (2019). The use uf classification trees to reduce selection bias for a non-probability sample with help from a probability sample. In Proceedings of the Survey Methods Section: SSC Annual Meeting, Calgary, AB, Canada.

Dever, J. (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. In Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference 1-15.

Deville, J.C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association* 87(418), 376–382.

Disogra, C., Cobb, C.L., Chan, E.K. and Dennis, J.M. (2011). Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics. *In Proceedings of the American Statistical Association, Section on Survey Research. Joint Statistical Meetings (JSM)*.

Elliot, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice* 2(6), 2982.

Elliott, M.R. (2022). Comments on "Statistical inference with non-probability survey samples".. Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00004-eng.htm.

Elliott, M. and Haviland, A. (2007). Use of a Web-Based Convenience Sample to Supplement a Probability Sample. *Survey Methodology* 33, 211-215.

Epanechnikov, V.A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* 14(1), 153-158.

Fahimi, M., Barlas, F.M., Thomas, R.K. and Buttermore, N. (2015). Scientific surveys based on incomplete sampling frames and high rates of nonresponse. *Survey Practice* 8(5), 1-11.

Ferri-García, R. and Rueda, M.M. (2018). Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys. SORT-Stat. Oper. Res. T., 42(2), 159–162.

Ferri-García, R. and Rueda, M.D.M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. PloS One, 15(4), e0231500.

Ferri-García, R., Beaumont, J.F., Bosa, K., Charlebois, J. and Chu, K. (2022). Weight smoothing for nonprobability surveys. *TEST* 31(3), 619-643.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics* 29, 1189-1232.

Hartley, H.O. (1962). Multiple frame surveys. *In Proceedings of the Social Statistics Section, American Statistical Association* 203-206.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association* 77(377), 89-96.

Kennedy, C. and Hartig, H. (2019). Response rates in telephone surveys have resumed their decline. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/

Kern, C., Li, Y. and Wang, L. (2021). Boosted Kernel Weighting - Using Statistical Learning to Improve Inference from Nonprobability Samples. *Journal of Survey Statistics and Methodology* 9(5), 1088-111.

Kim, J.K. and Tam, S.M. (2021). Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference. *International Statistical Review* 89, (2), 382401.

Kim, J.K. and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review* 87, 177-191.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. and Hunt, T. (2022). *caret: Classification and Regression Training.* https://CRAN.R-project.org/package=caret

Lee, B.K., Lessler, J. and Stuart, E.A. (2010). Improving propensity score weighting using machine learning. *Stat. Med.* 29, 337-346.

Lee, B.K., Lessler, J. and Stuart, E.A. (2011). Weight trimming and propensity score weighting. *PLoS ONE 6* e18174.

Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *J. Off. Stat.* 22(2), 329-349

Lee, S. and Valliant, R. (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociol. Method. Res.* 37(3), 319-343.

Liu, Z. and Valliant, R. (2023). Investigating an alternative for estimation from a non-probability sample: Matching plus calibration. *Journal of Official Statistics* 39, 45-78.

Lohr, S. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology* 37(2), 197-213.

Lohr, S.L. (2022). Comments on "Statistical inference with non-probability survey samples". Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00005-eng.htm.

Marken, S. (2018). Still Listening: The State of Telephone Surveys. Gallup Methodology Blog. https://news.gallup.com/opinion/methodology/225143/listening-state-telephone-surveys.aspx.

McCaffrey, D.F., Griffin, B.A., Almirall, D., Slaughter, M.E., Ramchand, R. and Burgette, L.F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine* 32(19), 3388-3414.

McCaffrey, D.F., Ridgeway, G. and Morral, A.R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods* 9(4), 403.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey methodology* 33(2), 151-157.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.F., Dessertaine, A. and Puech, P. (2022). QR Prediction for Statistical Data Integration. *TSE Working Paper* 22, 13-44.

Meng, X.-L. (2022). Comments on "Statistical inference with non-probability survey samples" - Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples. Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00006-eng.htm.

Nekrašaitė-Liegė, V., Čiginas, A. and Krapavickaitė, D. (2022). Usage of non-probability sample and scraped data to estimate proportions. *Workshop on survey statistics 2022.* https://vb.vgtu.lt/object/elaba:138918140/index.html

Rao, J.N.K.(2020). On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya B* 83(1), 242-272.

Rafei, A., Elliott, M.R. and Flannagan, C.A. (2022). Robust and efficient bayesian inference for non-probability samples. arXiv preprint arXiv:2203.14355.

Rivers, D. (2007). Sampling for web surveys. *In Proceedings of the 2007 Joint Statistical Meetings, Salt Lake City, UT, USA*.

Robbins, M.W., Ghosh-Dastidar, B. and Ramchand, R. (2021). Blending probability and nonprobability samples with applications to a survey of military caregivers. *Journal of Survey Statistics and Methodology* 9(5), 1114-1145.

Rueda, M.D.M., Ferri-García, R. and Castro-Martín, L. (2022). Combining Big Data with probability survey data: a comparison of methodologies for estimation from non-probability surveys. *Padua Research Archive-Institutional Repository* 711.

Rueda, M.D.M., Pasadas-del-Amo, S., Cobo, B., Castro-Martín, L. and Ferri-García, R. (2022). Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain. *Biometrical Journal*. https://doi.org/10.1002/bimj.202200035

Särndal, C.E. and Wright, R.L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics* 146-156.

Servy, E., Cuesta, C.B., Marí, G.P.D. and Armida, M.L. (2006). Utilización del paquete "kernsmooth" de r para construir suavizados loess y bandas de variabilidad a datos de la encuesta de ocupación hotelera. http://hdl.handle.net/2133/8792

Tillé, Y. and Matei, A. (2021). *sampling: Survey Sampling.* https://CRAN.R-project.org/package=sampling

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology* 8(2), 231-263.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite population sampling and inference: A prediction approach*. New York: John Wiley.

Wang, L., Graubard, B.I., Katki, H.A. and Li, Y. (2020). Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, https://doi.org/10.1111/rssa.12564.

Wang, Z. and Kim, J.K. (2022). Comments on "Statistical inference with non-probability survey samples". Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00007-eng.htm.

Wiśniowski, A., Sakshaug, J.W., Perez Ruiz, D.A. and Blom, A.G. (2020). Integrating Probability and Nonprobability Samples for Survey Inference. *Journal of Survey Statistics and Methodology* 8(1), 120-147. https://doi.org/10.1093/jssam/smz051.

Wolter, K. (2007). *Introduction to variance estimation.* New York: Springer.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, Statistics Canada, 48 (2). Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00002-eng.htm.

Xi, W., Hinton, A., Lu, B., Krotki, K., Keller-Hamilton, B., Ferketich, A. and Sukasih, A. (2022). Analysis of combined probability and nonprobability samples: a simulation evaluation and application to a teen smoking behaviour survey, *Communications in Statistics - Simulation and Computation* DOI: 10.1080/03610918.2022.2102181.

Yang, S. and Kim, J.K. (2020). Statistical data integration in survey sampling: a review. *Jpn J Stat Data Sci* 3, 625-650. https://doi.org/10.1007/s42081-020-00093-w

Yang, S., Kim, J. and Hwang, Y. (2021). Integration of survey data and big observational data for finite population inference using mass imputation. *Survey Methodology* 47, 29-58

A. Appendix 1

Regularity conditions for the HT estimator

The first and second order probabilities verify:

1a) $N^{-2}\sum_{i\neq j=1}^{N}(\pi_i\pi_j-\pi_{ij})^r = O(n^{-2r\delta})$

2a) $N^{-1}\sum_{i=1}^{N}(y_i/\pi_i - Y/n)^{2k} < M < \infty$ for $\delta > 0$ and $r^{-1}+k^{-1}=1$

Regularity conditions for the KW estimator:

The kernel function $K(u)$, the bandwidth h and the sampling schemes verify:

2a) $K(u)$, $\int K(u)du = 1$, $\sup_u |K(u)| < \infty$, y $\lim_{|u|\to\infty}|u||K(u)| = 0$

2b) $h = h(n_v)$, $h \to 0$, but $n_v h \to \infty$ as $n_v \to \infty$ and the distributions of the estimated propensity scores in the probability and non-probability samples are interchangeables.

SORT 48 (1) January-June 2024, 125-152 DOI: 10.57645/20.8080.02.16

Small area estimation of the proportion of single-person households: Application to the Spanish Household Budget Survey

María Bugallo Porto[1,*], Domingo Morales González[1]
and María Dolores Esteban Lefler[1]

Abstract

Household composition reveals vital aspects of the socioeconomic situation and major changes in developed countries for decision-making and mapping the distribution of single-person households is highly relevant and useful. Driven by the Spanish Household Budget Survey data, we propose a new statistical methodology for small area estimation of proportions and total counts of single-person households. Estimation domains are defined as crosses of province, sex and age group of the main breadwinner of the household. Predictors are based on area-level zero-inflated Poisson mixed models. Model parameters are estimated by maximum likelihood and mean squared errors by parametric bootstrap. Several simulation experiments are carried out to empirically investigate the properties of these estimators and predictors. Finally, the paper concludes with an application to real data from 2016.

MSC: *62J12, 62P25, 62D05.*

Keywords: *Small area estimation, zero-inflated Poisson mixed model, area-level data, Household Budget Survey, single-person household.*

1. Introduction

National statistical offices plan surveys to provide a cost effective way of obtaining accurate estimates at a certain level of aggregation. Nonetheless, disaggregated statistics can facilitate more effective targeting of decision-making, but obviously require more information to adequately represent population subgroups. If domain sample sizes are

* *Corresponding author:* mbugallo@umh.es

[1] Operations Research Center, Miguel Hernández University of Elche (Spain). Address: Edificio Torretamarit - Avda. de la Universidad s/n, 03202 Elche (Alicante).

Received: January 2023
Accepted: November 2023

large enough, we can accurately estimate domain characteristics using direct estimators, such as the Hájek estimator (Hájek, 1971). The term "small areas" is commonly used to describe domains with too small sample sizes to obtain precise direct estimates. In these cases, indirect estimation techniques, relying on statistical modelling, will have to be used. Small area estimation (SAE) addresses this challenge by borrowing strength from auxiliary variables, data from other domains and underlying dependency structures.

Given the topic of our research, in recent decades, most developed countries have faced major demographic changes that directly affect household composition (Cohen, 2021), with new forms of cohabitation replacing the traditional concept of "*two-parent family with children*" (Lesthaeghe, 2014). In a context of social transformation, living alone has become a sign of individual autonomy and freedom (Fritsch, Riederer and Seewann, 2023), even if it is sometimes still stereotyped (Greitemeyer, 2009). Meanwhile, loneliness and its impact on physical and mental health are an increasingly widespread problem (Snell, 2017), accentuating the symptoms of cognitive diseases (Lee and Lee, 2021; Park et al., 2016). Especially, among elderly single-person households, the need for medical care is expected to be high, and even more so compared to other age groups. Hence the natural need for research aimed at curbing these problems.

Among the main indicators of loneliness, we can mention the proportion and total count of single-person households by domains defined by territorial and demographic features. Indeed, the disaggregated mapping of these indicators provides valuable information for governments to implement social and health policies aimed at improving the well-being of people suffering from loneliness. Hence, more specific studies are needed. In addition, the number and size of households in the coming years is closely related to demographic projections (Ortiz-Ospina, 2019) and their distribution across provinces, sex and age groups is therefore of particular interest (Cho et al., 2019). For that purpose, this paper develops a new statistical methodology and illustrates its use with an application to the Spanish Household Budget Survey (SHBS), where the aim is to estimate proportions of single-person households by Spanish province, sex and age group. However, it can be applied to other contexts where the same problem holds.

The following is an overview of the state of the art. SAE uses linear mixed models (LMMs) and generalized linear mixed models (GLMMs) that can be fitted to either unit or area-level data. Area-level models have the advantage of easily incorporate auxiliary variables from statistical sources other than the sample. Namely, Torabi and Rao (2014) and Cai and Rao (2022) use subarea-models to deal with hierarchically structured data. Zhang and Chambers (2004) develops log-linear structural models suitable to estimate small area cross-classified counts based on survey data. Esteban et al. (2012), Marhuenda, Molina and Morales (2013); Marhuenda, Morales and Pardo (2014) and Morales, Pagliarella and Salvatore (2015) estimate poverty proportions based on LMMs. For GLMMs, binomial and multinomial mixed models are applied to estimate proportions by Molina, Saei and Lombardía (2007), Ghosh et al. (2009), Chandra and Chambers (2011), Chen and Lahiri (2012), Chambers, Salvati and Tzavidis (2012), López-Vizcaíno, Lombardía and Morales (2013, 2015), Militino, Ugarte and Goicoa

(2015), Chambers, Salvati and Tzavidis (2016), Hobza and Morales (2016), Liu and Lahiri (2017), as well as Hobza, Morales and Santamaría (2018). Poisson (PO) and Negative Binomial (NB) mixed models are employed to estimate counts and proportions by Dreassi, Petrucci and Rocco (2014), Tzavidis et al. (2015), Boubeta, Lombardía and Morales (2016, 2017) and Morales, Krause and Burgard (2022), among others. As for the computational limitations of PO-GLMMs, but with a unit-level approach, the conjugate form of the Gamma-PO model allows for computationally light estimation and prediction procedures (Berg, 2022). However, none of the above cited papers deal with data with excess zeros.

In scientific and technical studies it is common to find count data with many zeros (Zuur et al., 2009; Michael and Thomas, 2016). This is the case for our target variable, the count of single-person households by domains. A possible solution is to fit a Fay-Herriot (FH) model after a transformation, and apply the methodology of Berg and Fuller (2012) to obtain a non-zero variance estimate if the observed value is zero. Another approach is to consider models in which the probability of the null count is modified with respect to that which would correspond to a given probability distribution. Because of their flexibility, zero-inflated models play a relevant role. Without wishing to be exhaustive, we cite some papers where these models are used in SAE. Pfeffermann, Terryn and Moura (2008) consider situations where the value of the target variable is zero or an observation from a continuous distribution. They analyse the assessment of literacy proficiency with the possible outcome of zero, indicating illiteracy, or a positive score measuring the literacy level. Chandra, Bathla and Sud (2010) and Chandra and Sud (2012) introduce unit-level mixtures between zero and a LMM. They estimate domain means of continuous variables when the census vector contains a substantial proportion of zeros. Chandra and Chambers (2011) generalize their previous proposal by modelling logarithms. Anggreyani, Indahwati and Kurnia (2015) estimate infant mortality using plug-in predictors based on area-level mixed effects zero-inflated PO models. Krieg, Boonstra and Smeets (2016) and Sadik, Anisa and Aqmaliyah (2019) have carried out simulation experiments for unit-level mixtures between zero and a nested error regression model under a Bayesian approach. Hartono, Kurnia and Indahwati (2017) deal with area-level zero-inflated binomial models, with an application to unemployment data in Indonesia. Datta and Mandal (2015) and Sugasawa, Kubokawa and Ogasawara (2017) propose uncertain random effects, which are expressed as mixtures of a normal distribution and a one-point-at-zero distribution. Bugallo et al. (2023) model the number of fires in small areas using a zero-inflated NB mixed model.

Currently, there are no published studies that address the estimation of proportions of single-person households in small areas. However, it is essential for a more accurate implementation of social policies, as well as for clarifying certain economic aspects related to the housing sector and the private consumption of basic resources. Because of this challenge, we introduce a zero-inflated PO mixed model, that is, a mixture model with a logistic mixed model on a latent variable that indicates whether we count zero or according to a PO mixed model. Based on that model, we construct predictors of domain-

level counts and proportions. To estimate mean squared errors (MSE) of small area predictors, we lay out a parametric bootstrap method following González-Manteiga et al. (2007) and González-Manteiga et al. (2008). For assessment and illustration, several simulation experiments and a detailed application to real data are included. Regarding the latter, special attention has been reserved for the excess of zeros and the conciliation between the area-level model-based approach and the traditional survey sampling design-based approach. Further comparisons are also made with a FH model and an area-level zero-inflated NB mixed model.

The main document is organized as follows. Section 2 describes the data and SAE problem. Section 3 introduces the area-level zero-inflated PO mixed model. Section 4 provides model-based predictors of domain counts and proportions. Section 5 presents bootstrap-based confidence intervals (CI) of model parameters and MSE estimators of the predictors. Section 6 addresses the case study. Section 7 summarizes some conclusions. The paper includes supplementary material organized in four appendices. Appendix A describes the Laplace approximation to the model log-likelihood and the algorithm to calculate the maximum likelihood (ML) estimators of model parameters and obtain modal predictors of random effects. Appendix B empirically investigates the behaviour of the fitting algorithm, predictors and MSE estimators. Appendix C gives some additional simulation results. Appendix D maps relative root mean squared error (RRMSE) estimates for the application to real data.

2. Data and problem of interest

This paper presents and applies a new SAE methodology, based on an area-level zero-inflated PO mixed model, to estimate proportions of single-person households in small areas. The script has been approached from an applied point of view, in order to provide a reference text for future research on zero-inflated data in SAE. As far as the dataset is concerned, we use the 2016 SHBS (SHBS2016). The anonymized data file can be downloaded from the Spanish Statistical Office (INE) website. Regarding sample sizes, the SHBS2016 is designed to calculate precise direct estimators at NUTS 2 level, but it does not publish results at a lower level of aggregation. Below that level, sample sizes are quite small and direct estimators lose precision. In our research, we consider $D = 416$ domains defined at NUTS 3 level by Spanish province ($I = 52$) crossed by sex ($J = 2$) and age group ($K = 4$). Given the sample sizes of SHBS2016, we are faced with an SAE problem. In fact, the quartiles of the small area sample sizes are $q_0 = 1$, $q_{0.25} = 17$, $q_{0.5} = 34$, $q_{0.75} = 72$ and $q_1 = 367$, respectively. Therefore, it is desirable to use more sophisticated prediction methods rather than direct estimators. In terms of methodology, Section 2.1 describes our research framework; Section 2.2 introduces the explanatory variables of the case study and Section 2.3 focuses on the zero inflation problem.

2.1. Count and size variables

Further notation is introduced below. Formally, the finite population of Spanish households, U, can be partitioned in subpopulations U_{ijk}, $i \in \mathbb{I} = \{1,\ldots,I\}$, $j \in \mathbb{J} = \{1,\ldots,J\}$, $k \in \mathbb{K} = \{1,\ldots,K\}$, defined by province, sex (*sex1*: men, *sex2*: women) and age group (*age1*: less than 45 years; *age2*: between 46 and 55 years; *age3*: between 56 and 64 years; *age4*: 65 years or older) of the main breadwinner. This is to say, each U_{ijk} is disjoint and $U = \bigcup_{i=1}^{I}\bigcup_{j=1}^{J}\bigcup_{k=1}^{K} U_{ijk}$. Let N and N_{ijk} be the sizes of populations U and U_{ijk}, respectively.

At unit-level, the variable of interest is dichotomic, i.e. $y_{ijkl} = 1$ if the household $u_{ijkl} \in U_{ijk}$ is single-person and $y_{ijkl} = 0$, otherwise. Let $s = \bigcup_{i=1}^{I}\bigcup_{j=1}^{J}\bigcup_{k=1}^{K} s_{ijk}$ be a SHBS sample extracted from U, so that $s_{ijk} \subset U_{ijk}$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$. Let n and n_{ijk} be the sample sizes of s and s_{ijk}, respectively. For ease of exposition, we write $l = 1,\ldots,n_{ijk}$ for the households in s_{ijk} and $l = n_{ijk}+1,\ldots,N_{ijk}$ for the households in $U_{ijk}\backslash s_{ijk}$.

The domain parameters of interest are the total count and proportion of single-person households in U_{ijk}, i.e.

$$Y_{ijk} = \sum_{l=1}^{N_{ijk}} y_{ijkl}, \quad \overline{Y}_{ijk} = \frac{Y_{ijk}}{N_{ijk}}, \quad i \in \mathbb{I},\ j \in \mathbb{J},\ k \in \mathbb{K}. \tag{2.1}$$

Let w_{ijkl} be the household sampling weight of $u_{ijkl} \in U_{ijk}$. The sample count and the Hájek estimator of Y_{ijk} and N_{ijk} are

$$y_{ijk\cdot} = \sum_{l=1}^{n_{ijk}} y_{ijkl}, \quad \hat{Y}_{ijk}^{dir} = \sum_{l=1}^{n_{ijk}} w_{ijkl}\, y_{ijkl}, \quad \hat{N}_{ijk}^{dir} = \sum_{l=1}^{n_{ijk}} w_{ijkl}, \quad i \in \mathbb{I},\ j \in \mathbb{J},\ k \in \mathbb{K}.$$

The sample proportion and the Hájek estimator of $\overline{Y}_{ijk}$ are

$$\overline{y}_{ijk\cdot} = \frac{y_{ijk\cdot}}{n_{ijk}}, \quad \hat{\overline{Y}}_{ijk}^{dir} = \frac{\hat{Y}_{ijk}^{dir}}{\hat{N}_{ijk}^{dir}}, \quad i \in \mathbb{I},\ j \in \mathbb{J},\ k \in \mathbb{K}. \tag{2.2}$$

Once the count and size variables have been presented, it is important to be aware of the following scheme. Section 3 details the area-level zero-inflated PO mixed model and Section 4 proposes model-based predictors of the domain parameters defined in (2.1). Nevertheless, this requires an external file with auxiliary variables aggregated at domain level. In any case, it must contain the dependent variable of the area-level model, y_{ijk}, the size variable (offset), m_{ijk}, and a vector of domain-level auxiliary variables, $\boldsymbol{x}_{ijk}$ (see Section 2.2). As far as y_{ijk} and m_{ijk} are concerned, two options can be considered:
Option 1. Take $y_{ijk} = \lfloor \hat{Y}_{ijk}^{dir} \rfloor$ and $m_{ijk} = \lfloor N_{ijk} \rfloor$, where $\lfloor\cdot\rfloor$ is the closest integer operator. Let $\hat{\mu}_{yijk}$ be a model-based predictor of the expected value of y_{ijk}. The predictors of $\overline{Y}_{ijk}$ and Y_{ijk} are

$$\hat{\overline{Y}}_{ijk} = \frac{\hat{\mu}_{yijk}}{m_{ijk}}, \quad \hat{Y}_{ijk} = \hat{\mu}_{yijk}, \quad i \in \mathbb{I},\ j \in \mathbb{J},\ k \in \mathbb{K}.$$

By taking the direct estimators of domain totals as the dependent variable of the area-level model, Option 1 reconciles the area-level model-based approach and the sample design approach to inference in finite populations. This is an important argument in favour of Option 1. On the other hand, the fitting algorithm or the calculation of predictors may become unstable when the values of the dependent variable are large, which require more refined programming.

Option 2. Take $y_{ijk} = y_{ijk\cdot}$ and $m_{ijk} = n_{ijk}$. Let $\hat{\mu}_{yijk}$ be a model-based predictor of the expected value of y_{ijk}. The predictors of $\overline{Y}_{ijk}$ and Y_{ijk} are

$$\hat{\overline{Y}}_{ijk} = \frac{\hat{\mu}_{yijk}}{m_{ijk}}, \quad \hat{Y}_{ijk} = \hat{N}_{ijk}^{dir}\hat{\overline{Y}}_{ijk}, \quad i \in \mathbb{I},\, j \in \mathbb{J},\, k \in \mathbb{K}.$$

Boubeta, Lombardía and Morales (2016) applies Option 2 for area-level PO mixed models, as it is computationally more robust, but it does not include the sampling weights into the model. As omission of sampling weights is an important problem with Option 2, because it can lead to biased predictors, our choice of Option 1 is properly justified, even if it makes programming more difficult.

2.2. Domain-level auxiliary information

Population sizes and domain-level auxiliary variables have been estimated from the 2016 Spanish Labour Force Survey (SLFS). The SLFS is published quarterly, includes nearly 65,000 dwellings, equivalent to approximately 160,000 people, and collects data on the labour force and its various categories, as well as on the population outside the labour market. The anonymized data file can be downloaded from the INE website. The sample size of each quarterly SLFS is larger than three times the size of an annual SHBS. From the first to the last quarter of 2016, there are about $4 \cdot 160,000$ respondents. As there are $D = 416$ estimation domains, it is expected an average of 1538 respondents per domain. In order to improve our results, we jointly use data from the four quarters of 2016 and apply (2.2). In this way the effects of the variances of the covariate means on the properties of the prediction procedure are considered negligible.

The set of domain-level auxiliary variables is calculated by estimating the proportion of people in the following factor categories: *Citizenship*: Spanish (cit1) and foreign (cit2); *Education*: primary or less (edu1), basic secondary education (edu2), advanced secondary education (edu3) and higher education, such as university (edu4); *Labour situation*: employed (lab1), unemployed (lab2) and inactive (lab3); *Civil status*: unmarried (civ1), married (civ2), widower (civ3) and separated or divorced (civ4); *Dwelling mobility*: more than a year in the same dwelling (dwe1) and the opposite (dwe2). The above-mentioned auxiliary variables are proportions, bounded in the interval [0, 1], i.e. they are continuous variables, not binary indicators. Since the sum of proportions in the categories of each factor is one, and based on their socio-economic meaning, we omit one category from each factor. Namely, we have deleted cit2, edu2, lab3, civ1, dwe2.

2.3. Zero inflation

So far we have discussed the necessity of auxiliary information, but we have not addressed the problem of excess zeros, nor even demonstrated its occurrence. Nevertheless, it is important to assess the presence of false zeros to model counts of single-person households by province, sex and age group. The reason lies in the low number of respondents at some crosses and thus the difficulty of detecting single-person households. Throughout the paper, it will be shown why the incorporation of zero-inflated structures is more appropriate for the case study. As we assume that y_{ijk} counts the number of single-person households in U_{ijk}, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, it can be described by an area-level PO mixed model with offset parameter m_{ijk} and some explanatory variables. However, the target variable is aggregated by province, sex and age group, so that the number of households in s_{ijk} may be too small. Moreover, there are 28 domains with zero single-person households in SHBS2016. As the number of zeros seems to be too large, it has been decided to fit an area-level zero-inflated PO mixed model to (y_{ijk}, m_{ijk}), $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$.

Table 2.1. *Distribution of domains where single-person households are not observed in SHBS2016, by sex and age group of the main breadwinner.*

	sex		
age group	*sex1*	*sex2*	Total
age1	3	8	11
age2	2	8	10
age3	2	2	4
age4	3	0	3
Total	10	18	28

Table 2.1 presents the distribution of zeros by sex and age group in SHBS2016. It is shown that the 28 zeros are mainly concentrated in certain sex-age group categories. In fact, it can be suggested that single-person households inhabited by young and middle-age women are likely to be more difficult to capture in the count, i.e. their expected proportion is lower. The opposite is true at older ages. In any case, the number of zeros appears to be too large for what would be expected under a PO distribution. This motivates that a zero-inflated PO mixed model will have a better performance. Section 6 and Appendix B analytically justify the importance of incorporate the zero-inflated structure, both in terms of significance and goodness-of-fit. The area-level PO mixed model and the area-level zero-inflated PO mixed model will be compared and the latter will be chosen because it will give better results.

In order to test the dependence between the count of zeros/non zeros and provinces, sex and age groups, we have applied the Pearson's Chi-Squared test in $2 \times I$, $2 \times J$ and $2 \times K$ contingency tables, calculating p-values by Monte Carlo (MC). As a result, p-values close to 0.06 are reached for province and age group as inputs, increasing to 0.18

for sex. Based on Table 2.1 and the results of the above tests, we have decided to consider only age-group randomness to model zero-inflated probabilities. Furthermore, applying the same tests to assess the dependence between the count of single-person households (less/greater than 1, 2 or 3) and provinces, sex and age groups, only the randomness of the age group is significant. Guided by the promise of finding a good, simple model, Section 3 presents our methodological proposal.

3. Area-level zero-inflated Poisson mixed model

This section describes the area-level zero-inflated PO mixed model proposed as a basis to derive small area predictors of the proportion of single-person households by domains. All mathematical steps are detailed, justifying the soundness of what is presented. The formulation of the model is given in an orderly fashion, followed by the description of the fitting algorithm in Appendix A, the ML-Laplace approximation. Although the model is proposed in a general form, it is adapted for application to real data where appropriate. In fact, the model description is based on the stratification used in the real data example in Section 6, because it is in these domains that the need to incorporate a zero-inflated structure to model the response variable has been assessed. Even so, it is easily adaptable to other situations involving the general zero inflation problem.

Let us consider a count variable y_{ijk} taking values on $\mathbb{N}\cup\{0\}$, $i\in\mathbb{I}$, $j\in\mathbb{J}$, $k\in\mathbb{K}$. Let $D=IJK$ be the total number of y-values. As a particular case, a country divided into provinces, sex and age groups can be modelled as follows. Let z_{ijk}, $\boldsymbol{x}_{1,ijk}=(x_{1,ijk1},\ldots,x_{1,ijkq_1})$ and $\boldsymbol{x}_{2,ijk}=(x_{2,ijk1},\ldots,x_{2,ijkq_2})$ be latent (non observable) variables and $1\times q_1$ and $1\times q_2$ row vectors containing area-level explanatory variables, respectively. Define the vectors and matrices $\boldsymbol{y}_{ij}=\underset{1\le k\le K}{\text{col}}(y_{ijk})$, $\boldsymbol{z}_{ij}=\underset{1\le k\le K}{\text{col}}(z_{ijk})$, $\boldsymbol{X}_{1,ij}=\underset{1\le k\le K}{\text{col}}(\boldsymbol{x}_{1,ijk})$. $\boldsymbol{X}_{2,ij}=\underset{1\le k\le K}{\text{col}}(\boldsymbol{x}_{2,ijk})$, $\boldsymbol{y}=\underset{1\le i\le I}{\text{col}}(\underset{1\le j\le J}{\text{col}}(\boldsymbol{y}_{ij}))$, $\boldsymbol{z}=\underset{1\le i\le I}{\text{col}}(\underset{1\le j\le J}{\text{col}}(\boldsymbol{z}_{ij}))$, $\boldsymbol{X}_1=\underset{1\le i\le I}{\text{col}}(\underset{1\le j\le J}{\text{col}}(\boldsymbol{X}_{1,ij}))$ and $\boldsymbol{X}_2=\underset{1\le i\le I}{\text{col}}(\underset{1\le j\le J}{\text{col}}(\boldsymbol{X}_{2,ij}))$. In order to understand how the data are stacked according to the col(.) operator, we rely on the application to the SHBS2016 data as a useful example. In this dataset, the $D=416$ domains are sorted by age group and, within each age group, the Spanish provinces are concatenated, first for males and then for females.

Let $u_{1,k}$, $u_{2,ijk}$ be independent $N(0,1)$ random effects, $\boldsymbol{u}_1=\underset{1\le k\le K}{\text{col}}(u_{1,k})\sim N_K(\boldsymbol{0},\boldsymbol{I})$, $\boldsymbol{u}_2=\underset{1\le i\le I}{\text{col}}(\underset{1\le j\le J}{\text{col}}(\underset{1\le k\le K}{\text{col}}(u_{2,ijk})))\sim N_{IJK}(\boldsymbol{0},\boldsymbol{I})$, $\boldsymbol{u}=(\boldsymbol{u}_1^\top,\boldsymbol{u}_2^\top)^\top$. The bivariate vector (y_{ijk},z_{ijk}) follow an area-level zero-inflated PO (aZIP13) mixed model if

$$z_{ijk}\overset{ind}{\sim}\text{BE}(p_{ijk}),\ P(y_{ijk}=0/z_{ijk}=1)=1,\ P(y_{ijk}=t/z_{ijk}=0)=\frac{e^{-\mu_{ijk}}\mu_{ijk}^t}{t!},\ t\in\{0\}\cup\mathbb{N},$$

where $0<p_{ijk}<1$, $\mu_{ijk}=m_{ijk}\lambda_{ijk}$, $m_{ijk}\in\mathbb{N}$ is known, $\lambda_{ijk}>0$ and p_{ijk} and λ_{ijk} depend on the explanatory variables $\boldsymbol{x}_{1,ijk}$ and $\boldsymbol{x}_{2,ijk}$, on the regression parameters $\boldsymbol{\beta}_1=(\beta_{11},\ldots,\beta_{1q_1})^\top$ and $\boldsymbol{\beta}_2=(\beta_{21},\ldots,\beta_{2q_2})^\top$, and on the standard deviation parameters $\phi_1>$

0 and $\phi_2 > 0$ by means of the link functions

$$\text{logit}(p_{ijk}) = \log \frac{p_{ijk}}{1-p_{ijk}} = \boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k} = \sum_{\ell=1}^{q_1} x_{1,ijk\ell}\beta_{1\ell} + \phi_1 u_{1,k}$$

$$\log(\lambda_{ijk}) = \boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk} = \sum_{\ell=1}^{q_2} x_{2,ijk\ell}\beta_{2\ell} + \phi_2 u_{2,ijk}, \quad i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}.$$

Inverting the above functions, it follows that

$$p_{ijk} = \frac{\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}}{1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}}, \quad \lambda_{ijk} = \exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}, \quad i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}. \tag{3.1}$$

In short, the proposed model is a mixture of two mixed submodels. First, the BE submodel drives the mixture and incorporates the information derived from the excess of zeros. Subsequently, the PO submodel deals with the modelling of count variables. To complete its definition, it is assumed that $(y_{ijk}, z_{ijk})^\intercal$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, are independent conditioned to $\boldsymbol{u}$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_2^\intercal, \phi_1, \phi_2)^\intercal$ be the vector of model parameters and define $\xi_{ijk} = I_{\{0\}}(y_{ijk})$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$. This is to say, $\xi_{ijk} = 1$ if $y_{ijk} = 0$ and $\xi_{ijk} = 1$, otherwise. It holds that

$$\begin{aligned}
&P(y_{ijk}|u_{1,k}, u_{2,ijk}; \boldsymbol{\theta}) = \xi_{ijk}\left[p_{ijk} + (1-p_{ijk})e^{-\mu_{ijk}}\right] + (1-\xi_{ijk})\left[(1-p_{ijk})\frac{e^{-\mu_{ijk}}\mu_{ijk}^{y_{ijk}}}{y_{ijk}!}\right] \\
&= \left(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}\right)^{-1} \\
&\cdot \left\{\xi_{ijk}\left[\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\} + \exp\left\{-m_{ijk}\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}\right\}\right]\right. \\
&+ (1-\xi_{ijk})\exp\left\{y_{ijk}(\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}) - m_{ijk}\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}\right. \\
&\left.\left. + y_{ijk}\log m_{ijk} - \log y_{ijk}!\right\}\right\}, \ i \in \mathbb{I}, \ j \in \mathbb{J}, \ k \in \mathbb{K}.
\end{aligned}$$

By the independence assumptions, we have that

$$P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} P(y_{ijk}|u_{1,k}, u_{2,ijk};\boldsymbol{\theta}).$$

Therefore, the likelihood function of the aZIP13 mixed model is

$$\begin{aligned}
P(\boldsymbol{y};\boldsymbol{\theta}) &= \int_{\mathbb{R}^{K(1+IJ)}} P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) f_u(\boldsymbol{u})\, d\boldsymbol{u} \qquad (3.2) \\
&= \prod_{k=1}^{K}\int_{\mathbb{R}^{1+IJ}} \Big(\prod_{i=1}^{I}\prod_{j=1}^{J} P(y_{ijk}|u_{1,k}, u_{2,ijk};\boldsymbol{\theta}) f_{N(0,1)}(u_{2,ijk})\, du_{2,ijk}\Big) f_{N(0,1)}(u_{1,k})\, du_{1,k},
\end{aligned}$$

and the respective log-likelihood function is

$$\ell(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{k=1}^{K} \log \int_{\mathbb{R}^{1+IJ}} \Big(\prod_{i=1}^{I} \prod_{j=1}^{J} P(y_{ijk}|u_{1,k}, u_{2,ijk}; \boldsymbol{\theta}) f_{N(0,1)}(u_{2,ijk})\, du_{2,ijk} \Big) f_{N(0,1)}(u_{1,k})\, du_{1,k}.$$

Given $\boldsymbol{y}$, the ML estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta};\boldsymbol{y})$, where $\Theta = \mathbb{R}^{q_1+q_2} \times \mathbb{R}_+^2$ and $\mathbb{R}_+ = (0,\infty)$. The expression of $\ell(\boldsymbol{\theta};\boldsymbol{y})$ contains integrals in $\mathbb{R}^{1+IJ}$. To maximize it, two functions can be applied sequentially. The first one would compute the integral on $\mathbb{R}^{1+IJ}$ and the second one would perform the maximization on $\boldsymbol{\theta}$. As this approach is not efficient, Appendix A describes the ML-Laplace approximation as an alternative maximization method.

4. Prediction of totals and proportions

Under the assumption that y_{ijk}, $i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}$, follows the proposed aZIP13 mixed model, this section is devoted to the development of new small area predictors. Typical of the literature, the inference is focused on the expected values

$$\mu_{yijk} \triangleq E[y_{ijk}|\boldsymbol{u}_{ijk}] = m_{ijk}(1-p_{ijk})\lambda_{ijk},\ i \in \mathbb{I},\ j \in \mathbb{J},\ k \in \mathbb{K}, \tag{4.1}$$

where $p_{ijk} = p_{ijk}(u_{1,k})$ and $\lambda_{ijk} = \lambda_{ijk}(u_{2,ijk})$ are defined in (3.1). In an orderly fashion, first the plug-in predictor is introduced. Subsequently, the best predictor and its empirical version are derived (see Molina, Saei and Lombardía (2007) for futher details). At the expense of the theoretical properties, simpler alternatives are finally proposed looking for a better computational performance. Under a scenario based on SHBS2016, they will be compared in simulation experiments in Appendix B so as to justify the application to real data in Section 6.

Firstly, by plugging ML estimators and modal predictors, the population-based quantities given by (4.1) can be predicted using the plug-in (IN) predictor, defined as

$$\hat{\mu}_{yijk}^{in} = m_{ijk}\big(1+\exp\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1\hat{u}_{1,k}\}\big)^{-1} \exp\{\boldsymbol{x}_{2,ijk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2\hat{u}_{2,ijk}\}.$$

Among the different predictors that can be mentioned, this is the simplest approach to understand and the easiest to calculate. Indeed, its ease of interpretation and calculation, as well as its computational performance and execution times, are unsurpassed. Nevertheless, there are other potentially competitive alternatives. Let us define $\boldsymbol{y}_k = \underset{1\le i\le I}{\text{col}}\big(\underset{1\le j\le J}{\text{col}}(y_{ijk})\big)$, $\boldsymbol{u}_{2,k} = \underset{1\le i\le I}{\text{col}}\big(\underset{1\le j\le J}{\text{col}}(u_{2,ijk})\big)$, $\boldsymbol{v}_k = (u_{1,k}, \boldsymbol{u}_{2,k}^{\intercal})^{\intercal}$. The best predictor (BP) of (4.1) is $\hat{\mu}_{yijk}^{bp}(\boldsymbol{\theta}) = m_{ijk}E[(1-p_{ijk})\lambda_{ijk}|\boldsymbol{y}_k]$. The conditional expectation $E_{ijk} = E[(1-p_{ijk})\lambda_{ijk}|\boldsymbol{y}_k]$ is

$$E_{ijk} = \frac{\int_{\mathbb{R}^{1+IJ}} \big(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}\big)^{-1} \exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\} P(\boldsymbol{y}_k|\boldsymbol{v}_k) f(\boldsymbol{v}_k)\, d\boldsymbol{v}_k}{\int_{\mathbb{R}^{1+IJ}} P(\boldsymbol{y}_k|\boldsymbol{v}_k) f(\boldsymbol{v}_k)\, d\boldsymbol{v}_k}.$$

Denote the numerator and denominator of E_{ijk} by $A_{ijk} = A_{ijk}(\boldsymbol{y}_k, \boldsymbol{\theta})$ and $B_k = B_k(\boldsymbol{y}_k, \boldsymbol{\theta})$, respectively. Define $\xi_{rtk} = I_{\{0\}}(y_{rtk})$, $r \in \mathbb{I}$, $t \in \mathbb{J}$, $k \in \mathbb{K}$. It holds that

$$A_{ijk} = \int_{\mathbb{R}^{1+IJ}} \frac{\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}}{1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}} \prod_{r=1}^{I}\prod_{t=1}^{J} \omega_{rtk} f_{N(0,1)}(u_{1,k}) f_{N(0,1)}(u_{2,rtk})\, du_{1,k}\, du_{2,rtk},$$

$$B_k = \int_{\mathbb{R}^{1+IJ}} \prod_{r=1}^{I}\prod_{t=1}^{J} \omega_{rtk} f_{N(0,1)}(u_{1,k}) f_{N(0,1)}(u_{2,rtk})\, du_{1,k}\, du_{2,rtk},$$

$$\omega_{rtk} = (1+\exp\{\boldsymbol{x}_{1,rtk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\})^{-1} \Bigg\{ \xi_{rtk} \Bigg[\exp\{\boldsymbol{x}_{1,rtk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\} + \exp\Big\{ -m_{rtk} \exp\{\boldsymbol{x}_{2,rtk}\boldsymbol{\beta}_2 + \phi_2 u_{2,rtk}\} \Big\} \Bigg] + (1-\xi_{rtk}) \exp\Big\{ y_{rtk}(\boldsymbol{x}_{2,rtk}\boldsymbol{\beta}_2 + \phi_2 u_{2,rtk}) - m_{rtk} \exp\{\boldsymbol{x}_{2,rtk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u_{2,rtk}\} + y_{rtk} \log m_{rtk} - \sum_{a=1}^{y_{rtk}} \log a \Big\} \Bigg\}.$$

The empirical best predictor (EBP) is $\hat{\mu}^{ebp}_{yijk} = \hat{\mu}^{bp}_{yijk}(\hat{\boldsymbol{\theta}})$ and can be calculated by a MC method using antithetic variables to reduce variability Hobza and Morales (2016). The outline is as follows:

1. Calculate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1^{\mathsf{T}}, \hat{\boldsymbol{\beta}}_2^{\mathsf{T}}, \hat{\phi}_1, \hat{\phi}_2)^{\mathsf{T}}$.
2. For $s = 1, \ldots, S$, generate $u^{(s)}_{1,k}$, $u^{(s)}_{2,rtk}$ i.i.d. $N(0,1)$, $u^{(S+s)}_{1,k} = -u^{(s)}_{1,k}$, $u^{(S+s)}_{2,rtk} = -u^{(s)}_{2,rtk}$.
3. Calculate $\hat{\mu}^{ebp}_{yijk} = m_{ijk}\hat{A}_{ijk}/\hat{B}_k$, where

$$\hat{A}_{ijk} = \frac{1}{2S}\sum_{s=1}^{2S} \frac{\exp\{\boldsymbol{x}_{2,ijk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u^{(s)}_{2,ijk}\}}{1+\exp\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1 u^{(s)}_{1,k}\}} \prod_{r=1}^{I}\prod_{t=1}^{J} \hat{\omega}_{rtk},\ \hat{B}_k = \frac{1}{2S}\sum_{s=1}^{2S}\prod_{r=1}^{I}\prod_{t=1}^{J} \hat{\omega}_{rtk}, \tag{4.2}$$

$$\hat{\omega}_{rtk} = \frac{1}{1+\exp\{\boldsymbol{x}_{1,rtk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1 u^{(s)}_{1,k}\}} \Bigg\{ \xi_{rtk} \Bigg[\exp\{\boldsymbol{x}_{1,rtk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1 u^{(s)}_{1,k}\} + \exp\Big\{ -m_{rtk} \exp\{\boldsymbol{x}_{2,rtk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u^{(s)}_{2,rtk}\} \Big\} \Bigg] + (1-\xi_{rtk}) \exp\Big\{ y_{rtk}(\boldsymbol{x}_{2,rtk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u^{(s)}_{2,rtk}) - m_{rtk} \exp\{\boldsymbol{x}_{2,rtk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u^{(s)}_{2,rtk}\} + y_{rtk} \log m_{rtk} - \sum_{a=1}^{y_{rtk}} \log a \Big\} \Bigg\},\ \xi_{rtk} = I_{\{0\}}(y_{rtk}).$$

It has been noted that (4.2) contains products with IJ terms. Given the nature of our problem, these products are close to zero under Option 1, leading to numerical precision problems in Section 6 and Appendix B. Facing this challenge, we have introduced a simplified version of the BP by conditioning to y_{ijk} instead of $\boldsymbol{y}_k$. This simplified predictor (SP) is $\hat{\mu}^{sp}_{yijk}(\boldsymbol{\theta}) = m_{ijk}E[(1-p_{ijk})\lambda_{ijk}|y_{ijk}]$. The conditional expectation

$E^{sp}_{ijk} = E[(1-p_{ijk})\lambda_{ijk}|y_{ijk}]$ is

$$E^{sp}_{ijk} = \frac{\int_{\mathbb{R}^2}\left(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1+\phi_1 u_{1,k}\}\right)^{-1}\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2+\phi_2 u_{2,ijk}\}P(y_{ijk}|\boldsymbol{u}_{ijk})f(\boldsymbol{u}_{ijk})\,d\boldsymbol{u}_{ijk}}{\int_{\mathbb{R}^2}P(y_{ijk}|\boldsymbol{u}_{ijk})f(\boldsymbol{u}_{ijk})\,d\boldsymbol{u}_{ijk}},$$

Denote the numerator and denominator of E^{sp}_{ijk} by $A^{sp}_{ijk} = A^{sp}_{ijk}(y_{ijk},\boldsymbol{\theta})$ and $B^{sp}_{ijk} = B^{sp}_{ijk}(y_{ijk},\boldsymbol{\theta})$, respectively. It holds that

$$A^{sp}_{ijk} = \int_{\mathbb{R}^2}\frac{\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2+\phi_2 u_{2,ijk}\}}{(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1+\phi_1 u_{1,k}\})}\,\omega_{ijk}\,f_{N(0,1)}(u_{1,k})f_{N(0,1)}(u_{2,rtk})\,du_{1,k}\,du_{2,rtk},$$

and

$$B^{sp}_{ijk} = \int_{\mathbb{R}^2}\omega_{ijk}\,f_{N(0,1)}(u_{1,k})f_{N(0,1)}(u_{2,rtk})\,du_{1,k}\,du_{2,rtk}.$$

The empirical simplified predictor (ESP) is $\hat{\mu}^{esp}_{yijk} = \hat{\mu}^{sp}_{yijk}(\hat{\boldsymbol{\theta}})$ and can be approximated by numerical approximation of integrals. However, the following antithetical MC algorithm is applied:

1. Calculate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^{\mathsf{T}}_1, \hat{\boldsymbol{\beta}}^{\mathsf{T}}_2, \hat{\phi}_1, \hat{\phi}_2)^{\mathsf{T}}$.
2. For $s = 1,\ldots,S$, generate $\boldsymbol{u}^{(s)}_{ij} = \left(u^{(s)}_{1,k}, u^{(s)}_{2,ijk}\right)^{\mathsf{T}}$ i.i.d. $N_2(\boldsymbol{0},\boldsymbol{I}_2)$, $\boldsymbol{u}^{(S+s)}_{ij} = -\boldsymbol{u}^{(s)}_{ij}$.
3. Calculate $\hat{\mu}^{esp}_{yijk} = m_{ijk}\hat{A}^{sp}_{ijk}/\hat{B}^{sp}_{ijk}$, where

$$\hat{A}^{sp}_{ijk} = \frac{1}{2S}\sum_{s=1}^{2S}\frac{\exp\{\boldsymbol{x}_{2,ijk}\hat{\boldsymbol{\beta}}_2+\hat{\phi}_2 u^{(s)}_{2,ijk}\}}{(1+\exp\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1+\hat{\phi}_1 u^{(s)}_{1,k}\})}\hat{\omega}_{ijk} \quad \text{and} \quad \hat{B}^{sp}_{ijk} = \frac{1}{2S}\sum_{s=1}^{2S}\hat{\omega}_{ijk}.$$

Because of the numerical precision of R, calculating exponential functions to predict μ_{yijk} may result in negative values that are too small. Consequently, ω_{ijk} would be close to zero. These overflow problems were detected by Boubeta, Lombardía and Morales (2016) and motivated these authors to choose Option 2, more computationally stable. In our case, as defined, the ESP allows us to solve them. Therefore, we assume Option 1, which is more convenient, as it reconciles to some extent the design-based and model-based approaches. Consequently, in simulations experiments in Appendix B and the case study in Section 6, the ESP will be used and the EBP will be omitted.

5. Bootstrap inference

This section presents bootstrap-based CIs for the model parameters and estimators of the MSEs of the predictors. For the latter, we adapt the procedures used by González-Manteiga et al. (2007, 2008).

5.1. Confidence intervals for model parameters

Let θ_ℓ be a component of the vector of model parameters $\boldsymbol{\theta}$. Let $\alpha \in (0,1)$. The following procedure calculates a $(1-\alpha)\%$ percentile bootstrap CI for θ_ℓ.

1. Fit the model to the sample and calculate the ML estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1^\intercal, \hat{\boldsymbol{\beta}}_2^\intercal, \hat{\phi}_1, \hat{\phi}_2)^\intercal$.
2. Repeat B times ($b = 1, \ldots, B$):
   (a) For $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, generate $u_{1,k}^{*(b)} \sim N(0,1)$, $u_{2,ijk}^{*(b)} \sim N(0,1)$ and calculate
   $$p_{ijk}^{*(b)} = \exp\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1 u_{1,k}^{*(b)}\} \left(1 + \exp\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1 u_{1,k}^{*(b)}\}\right)^{-1}, \quad (5.1)$$
   $$\lambda_{ijk}^{*(b)} = \exp\{\boldsymbol{x}_{2,ijk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u_{2,ijk}^{*(b)}\} \ .$$
   (b) Generate $z_{ijk}^{*(b)} \sim \mathrm{BE}(p_{ijk}^{*(b)})$. If $z_{ijk}^{*(b)} = 1$, do $y_{ijk}^{*(b)} = 0$. If $z_{ijk}^{*(b)} = 0$, generate $y_{ijk}^{*(b)} \sim \mathrm{PO}(m_{ijk}\lambda_{ijk}^{*(b)})$.
   (c) On the basis of the bootstrap sample $(y_{ijk}^{*(b)}, m_{ijk}, \boldsymbol{x}_{ijk})$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, calculate the ML estimate $\hat{\theta}_\ell^{*(b)}$.
3. Sort the values $\hat{\theta}_\ell^{*(b)}$, $b = 1, \ldots, B$, from smallest to largest. They are $\hat{\theta}_{\ell(1)}^* \leq \ldots \leq \hat{\theta}_{\ell(B)}^*$. A $(1-\alpha)\%$ percentile bootstrap CI for θ_ℓ is $\left(\hat{\theta}_{\ell(\lfloor(\alpha/2)B\rfloor)}^*, \hat{\theta}_{\ell(\lfloor(1-\alpha/2)B\rfloor)}^*\right)$.

5.2. Mean squared error estimation

The model-based MSE of the EBP, ESP or IN predictor, $\hat{\mu}_{yijk}$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, can be estimated using a resampling method. The following procedure calculates a parametric bootstrap estimator of $MSE(\hat{\mu}_{yijk})$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$.

1. Fit the model to the sample and calculate the ML estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1^\intercal, \hat{\boldsymbol{\beta}}_2^\intercal, \hat{\phi}_1, \hat{\phi}_2)^\intercal$.
2. Repeat B times ($b = 1, \ldots, B$):
   (a) Run Steps (a) and (b) of the algorithm detailed in Section 5.1.
   (b) For $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, calculate $\mu_{yijk}^{*(b)} = m_{ijk}(1 - p_{ijk}^{*(b)})\lambda_{ijk}^{*(b)}$.
   (c) On the basis of the bootstrap sample $(y_{ijk}^{*(b)}, m_{ijk}, \boldsymbol{x}_{ijk})$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, calculate the ML estimate $\hat{\boldsymbol{\theta}}^{*(b)}$ and the predictor $\hat{\mu}_{yijk}^{*(b)}$.
3. Output: $mse^*(\hat{\mu}_{yijk}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\mu}_{yijk}^{*(b)} - \mu_{yijk}^{*(b)}\right)^2$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$.
4. An estimator of the model-based MSE of the Hájek estimator is
$$mse^*(\hat{\bar{Y}}_{ijk}) = \frac{1}{B}\sum_{b=1}^{B}\left(y_{ijk}^{*(b)} - \mu_{yijk}^{*(b)}\right)^2, \quad i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}.$$

6. Application to real data

As a starting point, some considerations are presented to place the application in context and to encourage us in the work we are about to undertake. Regarding 2016, the Spanish Household Projection 2016–2031 addresses demographic trends and social patterns currently observed in Spain in terms of the number of households. Its authorship is attributed to the INE. It shows that households will increase by 4.9%, despite the decrease in the number of inhabitants, because of a reduction in the expected number of residents per dwelling, from 2.50 in 2016 to 2.35 in 2031. Related to this, between 2016 and 2031 the smallest households (one or two people in a shared dwelling) would continue to grow, while the largest ones would decrease, with a relative increase of 19.6% of single-person households. As a result, there will be more than 5.5 million single-person households (28.6%), with 12% of the Spanish population living alone.

For methodological purposes, this section applies the aZIP13 mixed model to the SHBS2016 data so as to estimate proportions of single-person households in small areas. Regarding the SHBS, it is published annually by the INE to study the nature and destination of consumer spending and the living conditions of households. The SHBS2016 includes around 22,000 dwellings, selected by means of a two-stage stratified random sampling carried out independently in each Autonomous Community (NUTS 2 level). Broadly speaking, the first stage units are territories with around 2,000 dwellings, called census sections. The second stage units are dwellings, interviewing all individuals over 16 years of age who reside in them. In each NUTS 2 region, the first stage units are stratified following a geographical criterion, which assigns the stratum according to the size of the municipality to which the section belongs. Sections are selected within each stratum with probability proportional to their population size. Dwellings are selected, within each section, with equal probability by means of systematic sampling with random start. The target variable y_{ijk} is the direct estimate of the number of single-person households in a domain where i, j and k represent the province of residence, sex and age group of the main breadwinner, respectively. Furthermore, direct estimates of population sizes and area-level auxiliary variables have been obtained from the four 2016SLFS microdata. What is more, they have been considered as true population values because of the precision derived from the acceptable sample sizes of the 2016SLFS surveys.

Table 6.1 shows the ML estimates of the regression parameters (RP) $\boldsymbol{\beta}_1$, ϕ_1 (BE submodel), $\boldsymbol{\beta}_2$ and ϕ_2 (PO submodel), the p-values to test $H_0 : \beta_{t\ell} = 0$, $t = 1,2$, $\ell = 1,...,q_t$, and $H_0 : \phi_t = 0$, $t = 1,2$, and the normal-asymptotic and bootstrap CIs at a 95% confidence level. For convenience, their lower (LB) and upper (UB) bounds are provided. Normal-asymptotic CIs are discussed in Appendix A and bootstrap CIs in Section 5.

The final model incorporates only those variables that are significant at 5%. The flexibility achieved by making the random effects of the count model domain-dependent allows us to reduce the importance of the set of domain-level variables and incorporate

only those that actually add relevant knowledge. In order, edu4, ci1, edu1, tm1, ec4, lab1, lab2 are removed. The BE submodel contains one auxiliary variable, $x_{1,1}$ = intercept, and the PO submodel four: $x_{2,1}$ = intercept, $x_{2,2}$ = edu3, $x_{2,3}$ = civ2 and $x_{2,4}$ = civ3.

The only parameter of the BE submodel, β_{11}, is significantly non-zero and its CIs have an acceptable short length, which guarantees some precision in its estimation. Actually, the latter provides strong evidence in favour of the zero-inflated structure. According to Table 6.1, none of the area-level auxiliary variables is relevant to explain the zero-inflated probabilities, supporting our contribution. Null counts are caused by the difficulty of detecting single-person households in domains with small sample sizes. The basic zero-inflated probability is $p_0(\hat{\beta}_{11}) = 0.063$, which implies that the basic probability of obtaining an observation from the PO submodel is 0.937. However, it has already been proven that it is also important to take into account the age-group randomness. Here, it is confirmed that the asymptotic and bootstrap 95% CIs for ϕ_1 do not contain the zero.

Table 6.1. *Regression parameters of the final aZIP13 mixed model.*

		BE submodel		PO submodel				
	RP	β_{11}	ϕ_1	β_{21}	β_{22}	β_{23}	β_{24}	ϕ_2
	Estimate	-2.696	0.398	-1.857	2.138	-0.649	3.881	0.517
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Asymp.	LB 95%	-3.270	0.091	-2.319	1.007	-1.057	3.207	0.482
	UB 95%	-2.121	1.752	-1.395	3.269	-0.242	4.554	0.555
Boot.	LB 95%	-3.317	0.0002	-2.312	1.051	-1.016	3.215	0.480
	UB 95%	-2.162	0.859	-1.432	3.270	-0.222	4.577	0.554

For the PO submodel, it could be suggested that a medium-high level of education (β_{22}), as well as being widower (β_{24}), contribute to increase the count of single-person households by domains, because their signs are significantly positive. On the other hand, an increase in the proportion of people who are married (β_{23}) implies a decrease in the number of single-person households, assuming that the other auxiliary variables are fixed. Given the group effect, it can be inferred that Spanish citizenship, employment status and dwelling mobility are not relevant to model the count of single-person households. The proportion of inhabitants with primary or university education and the proportion of separated or divorced people are also irrelevant. Last but not least, the asymptotic and bootstrap 95% CIs for ϕ_2 do not contain the zero, confirming the necessity of modelling the counts with a random-effect model.

Back to the modelling of the zero-inflated structure, recall that Table 2.1 suggested that the number of zeros appears to be too large for what would be expected under a PO distribution. This statement is confirmed by comparing the number of zeros found in $B =$ 1000 bootstrap resamples under the PO mixed model and the proposed aZIP13 mixed

model. Indeed, for the PO mixed model there are no null counts in any resample, with 38 single-person households in the lowest case. The number of zeros in the data exceeds the number of zeros that could plausibly be generated by the fitted PO distribution. For the aZIP13 mixed model, each resample contains an average of 28 zeros, closely mimicking the structure of Table 2.1 and, thus, the behaviour of the target variable.

Hereafter, we assume the aZIP13 mixed model that Table 6.1 presents. To have more confidence in this model as a true generating model, Section 6.1 addresses its validation and Appendix B performs some simulation experiments under the SHBS2016 scenario. Importantly, they support the use of the IN predictor.

6.1. Model validation

Residual analysis is used to validate a model as well as to detect potential underlying dependency relationships. As the aZIP13 mixed model is an area-level model, model diagnosis is also performed at that level of aggregation. Besides, we are interested in the conciliation of the model-based approach and the design-based approach to SAE. Further notation is introduced below. Let us define the raw residuals (RR) as $e_{ijk} = y_{ijk} - \mu^{in}_{yijk},\ i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}$. Under Option 1, $y_{ijk} = \lfloor \hat{Y}^{dir}_{ijk} \rfloor$ and $e_{ijk} = \lfloor \hat{Y}^{dir}_{ijk} \rfloor - \mu^{in}_{yijk}$. The standardized residuals (SR) are defined by dividing the RRs by its standard deviation.

In what follows, validation results are shown for a better interpretation of the application to real data. To start with, Figure 6.1 plots the SRs of the aZIP13 mixed model versus domain indexes (left) and predicted values of the proportion of single-person households in original (center) and log scale (right). In dotted red, the line $y = 0$ is added.

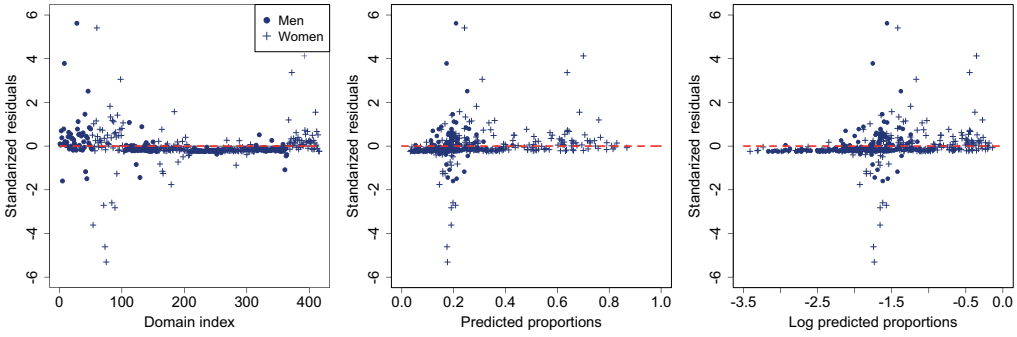


Figure 6.1. *SRs versus domain indexes (left) and predicted values of the proportion of single-person households in original (center) and log scale (right).*

As general conclusions drawn from Figure 6.1, it can be seen that SRs have a pattern of symmetry around zero and are mainly found in $[-3,3]$. The central plot has a low percentage of domains with large predicted probabilities, which exceed the threshold of 0.7, and correspond to domains with predominantly single-person households, i.e. inhabited by elderly women. Regarding the right plot, plotting SRs against log predicted

probabilities allows us to detect a conical pattern in the scatterplot. That is, as the log predicted probabilities increases, so does the variability of SRs. This phenomenon is in agreement with the theoretical dispersion of the aZIP13 mixed model.

As expected, SRs are highly variable between provinces, with different sample sizes and socioeconomic conditions. Namely, there are 11 areas with absolute SRs greater than 3, which represents 2.650% of the domains. Directly related to housing prices, the most affected are Madrid and Barcelona. Related to age group and sex, changes are minor. However, *age4* contains the largest amount of outliers.

6.1.1. Zero inflation validation

Area-level models do not aim to chase the scatterplot, but to smooth it and provided more accurate results. It is therefore crucial to understand the importance of zero-inflated probabilities, as they solve the problems of overfitting of the PO mixed model to the Hájek estimates. Indeed, Figure 6.2 shows this improvement in domains with null counts of single-person households (left) and with less than 5 counts (center). All observations are sorted according to the domain index. The line charts plot the Hájek estimates, the IN predictions and those relative to the IN predictor of the PO mixed model with the same set of area-level auxiliary variables as the aZIP13 mixed model, denoted as IN0. The advantage of the IN predictor over the IN0 predictor also applies when comparing with the IN predictor that uses a constant zero-inflated probability $p_{ijk} = p$, denoted as IN1 in Appendix B, although it is not included for ease of exposition.

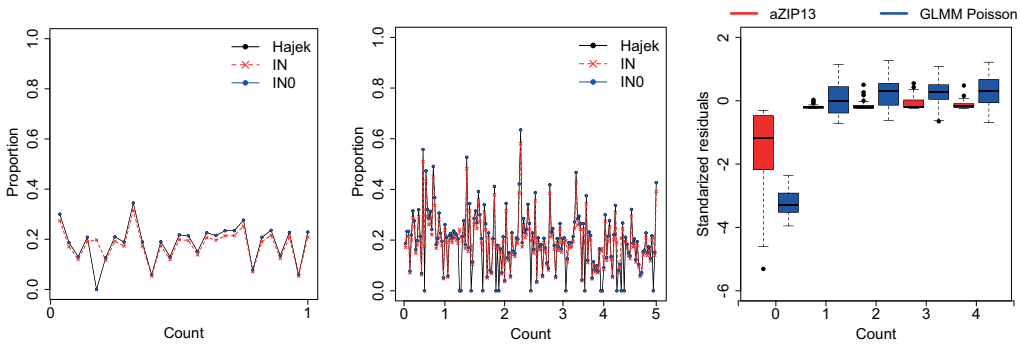


Figure 6.2. *Predicted proportions of single-person households in domains with null counts (left) and less than 5 counts (center) and boxplots for the respective SRs (right).*

On balance, the IN predictor of the aZIP13 mixed model is the one that smoothes the results the most. In addition, a challenge encountered in modelling direct estimators is that, in areas with tiny sample sizes, some households in the sample represent too many households in the population. The main concern is to find area-level outliers. Figure 6.2 (right) shows boxplots of the SRs from the aZIP13 mixed model and the PO mixed model in domains with less than 5 counts, grouped according to the observed single-person household counts. When single-person households are not observed, the

PO mixed model is clearly worse and, for low counts, its performance does not improve either: the variability of the boxes is higher. It is concluded that the aZIP13 mixed model performs satisfactorily, both in terms of the significance level of the RPs, the validation via SRs and the fit of zero outcomes. This is a great support for the proposed methodology.

6.2. Predictions and error measures

This section provides Hájek estimates and IN predictions of the proportion of single-person households by province, sex and age group. Figure 6.3 shows line charts of these values sorted by domain index (left) and sample size (center), as well as a comparison of both (right). Among the most noteworthy findings, model-based predictors correct the excessively large Hájek estimates, especially for elderly women in Madrid and Barcelona. Even more, it is inferred that the IN predictor smoothes the results of the Hájek estimator, although it still presents problems when dealing with extreme proportions. On the other hand, if single-person households are not observed, the Hájek estimator has no margin of error, although the model never comes to such a low proportion. The same is true for values close to one. This can be seen in Figure 6.3 (left). As it is unlikely, our research is a methodological improvement. In addition, it can be observed that household composition does not affect all domains equally: as the age group increases, the proportion of single-person households also increases.

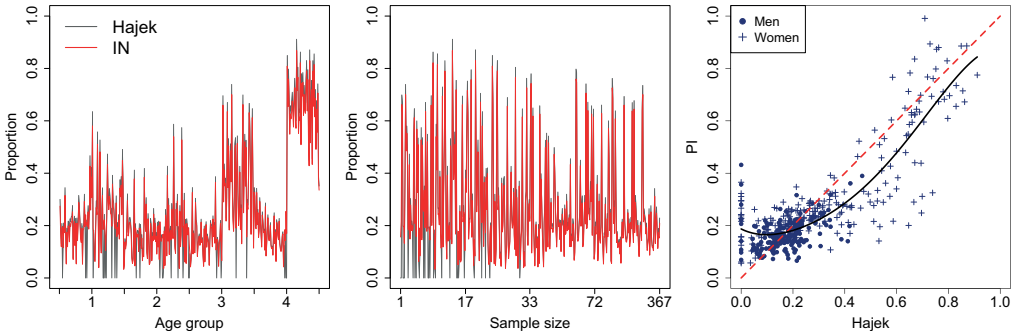


Figure 6.3. *IN proportions of single-person households sorted by domain (left) and sample size (center), and Hájek estimates versus IN proportions (right).*

According to Figure 6.3 (center), the IN predictor gets closer to the Hájek estimator as the sample size increases, which is one of the most convincing aspects of the data analysis. Eventually, Figure 6.3 (right) plots the Hájek estimates versus the IN proportions. It can be seen that the dots are evenly distributed around $y = x$. To support this statement, a local polynomial regression of degree 3, with an appropriate bandwidth, is plotted to smoothly represent the relationship between ordinates and abscissas. Consequently, we can underline a crucial advantage of our approach: the theoretical properties of the Hájek estimator, such as asymptotic design-based unbiasedness, are, to some extent, inherited by the IN predictor based on the aZIP13 mixed model.

Table 6.2 (a) reports IN proportions of single-person households by sex and age group. Predominantly, it is the population of *age2* that is least likely to live in single-person households, followed by *age1*. The current trend projects an increase in the proportion of single-person households, with the number of households inhabited by elderly women skyrocketing. This phenomenon is associated with the ageing process, which progressively involves the emancipation of children and widowhood.

Table 6.2. *Tabular results for the IN predictor and RRMSEs (%) of the proportion of single-person households by sex and age group of the main breadwinner.*

	sex		
age group	*sex1*	*sex2*	Total
age1	0.1988	0.2389	0.2187
age2	0.1596	0.1838	0.1736
age3	0.1468	0.3694	0.2612
age4	0.1707	0.6479	0.4394
Total	0.1830	0.3371	0.2621

(a) *IN proportions aggregated by province.*

	sex		
age group	*sex1*	*sex2*	Total
age1	20.8360	19.8250	20.3350
age2	20.3787	22.0940	21.2451
age3	20.9409	12.6800	16.6921
age4	20.1576	18.1149	19.0072
Total	20.6300	19.3125	19.9537

(b) *IN RRMSEs (%) aggregated by province.*

As for the error measures, we calculate the parametric bootstrap estimator of the MSE of μ^{in}_{yijk}, $i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}$, following Section 5. $B = 2000$ resamples are used. To avoid scale dependencies, and as usual, the script should be focused on RRMSEs. However, the non-relative version, the root-MSE (RMSE), is preferable because it allows a better understanding of what happens with null counts. Accordingly, Figure 6.4 plots model-based estimates of RMSEs for the IN predictor versus design-based standard deviations (RVAR) for the Hájek estimator (left) and versus model-based estimates of RMSEs for the Hájek estimator (right). See Morales et al. (2021) (Section 2.5) for further details about the RVARs of the Hajék estimator.

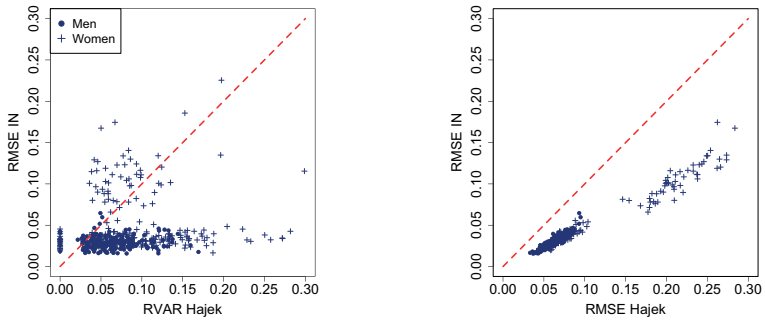


Figure 6.4. *Model-based estimates of RMSEs for the IN predictor versus design-based RVARs (left) and model-based estimates of RMSEs (right) for the Hájek estimator.*

Broadly speaking, both plots in Figure 6.4 show that the IN predictor has lower RMSE in all domains and, in most of them, it is also lower than the design-based RVAR of the Hájek estimator. The reduction of the model-based RMSE is therefore prominent when we use the IN predictor instead of the Hájek estimator. Nevertheless, the Hájek estimator has an estimated variance of zero for an observed zero and the RMSE of the IN predictor is always greater than zero. As there are 28 zeros in SHBS2016, this implies 28 aligned points in the lower left corner of Figure 6.4 (left). Clearly, we have already reported that these are false zeros.

In terms of magnitude, the RMSE is higher for elderly women, and it is attributable to the high predicted and/or estimated proportions for these domains. Therefore, it is also useful to provide summary measures of the RRMSE, expressing the error in percentage terms. Table 6.2 (b) contains the bootstrap estimates of the RRMSE (in %) for the IN predictor by sex and age group. As a general conclusion, all values are around 20%, with a slightly lower average for women and especially for *age3*. Hence, the IN predictions of the proposed aZIP13 mixed model have low RRMSEs, as expected in SAE. Appendix D of Supplementary Material maps these relative errors by province, sex and age group.

6.3. Mapping proportions of single-person households

The case study concludes by analysing the socioeconomic findings drawn from the area-level predictions. In this sense, the proposed methodology offers the opportunity to analytically read the appreciable differences by Spanish province, sex and age group. Figures 6.5–6.8 map the provincial distribution of single-person households for men (left) and women (right) according to the age group of the main breadwinner.

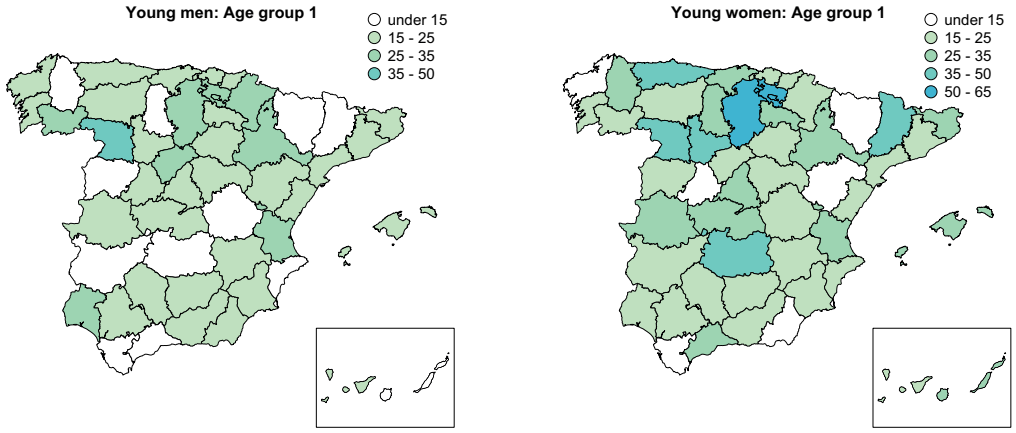


Figure 6.5. *Percentages of single-person households for young men (left) and women (right).*

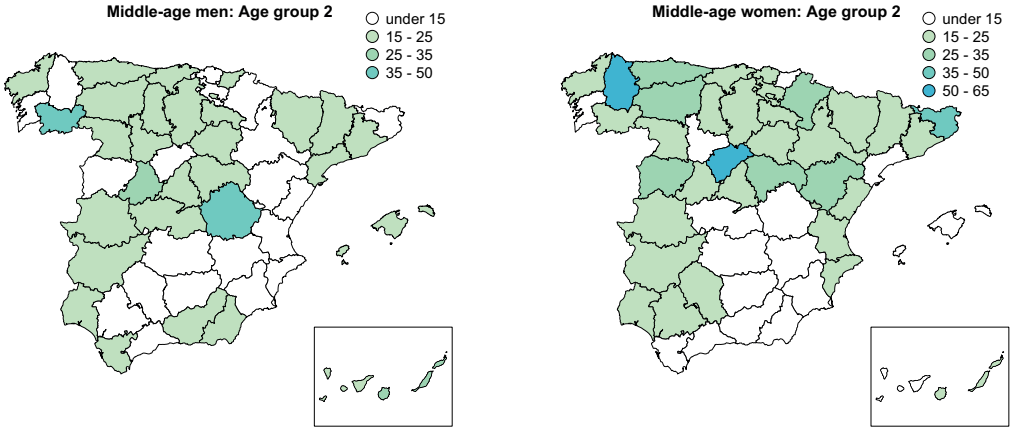


Figure 6.6. *Percentages of single-person households for middle-age men (left) and women (right).*

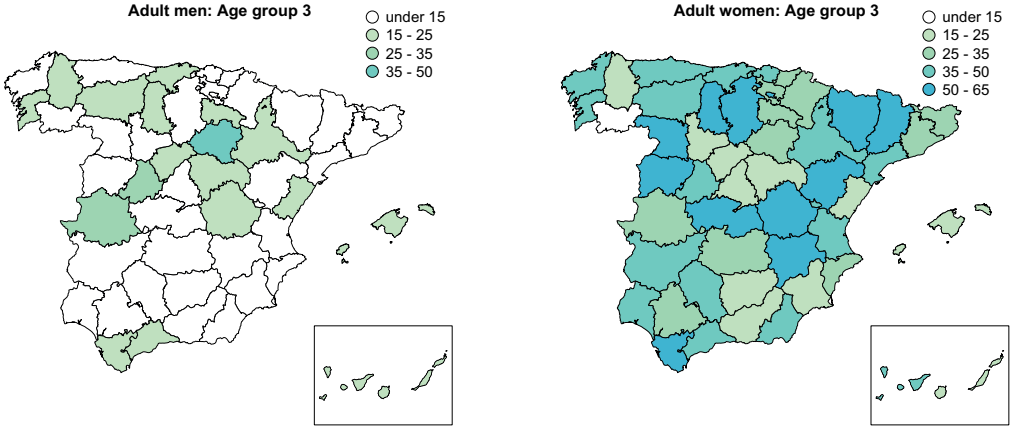


Figure 6.7. *Percentages of single-person households for adult men (left) and women (right).*

On the one hand, they show that the highest proportions of single-person households are found in the centre and north-west of Spain, with lower rates in the south and Canary Islands. As expected, the distribution between neighboring provinces, or between those whose demographic and socioeconomic conditions are similar, is generally homogeneous. This fact justifies how model-based predictors lead to smoother results (and closer to reality) than direct estimators. In addition, an interesting spatial pattern emerges, as it can be observed an inverse relationship between house prices and the proportion of single-person households. Thus, lower proportions are estimated for the Catalan Coast, Madrid, Balearic Islands and Málaga. In other words, the Spanish provinces with the highest average prices.

On the other hand, over the course of a person's life, their lifestyle can be expected to change, with the age group directly affecting the composition of households. Most

notably, old age is linked to another factor that alters household composition: mortality. So sex and *age4* are crucial here. Moreover, the increase in quality of life implies not only an increase in life expectancy but also in the autonomy of the elderly, which results in an increase in the number of single-person households inhabited by retired people. Most men live with their partners until their death. In contrast, women have a longer life expectancy (implying a greater accumulation at the top of the demographic pyramid) and the average age of their partners is higher, so they will live alone to a greater extent. Accordingly, Figures 6.5–6.8 map a significant difference between men and women, with clearly higher proportions of dwellings inhabited only by women.

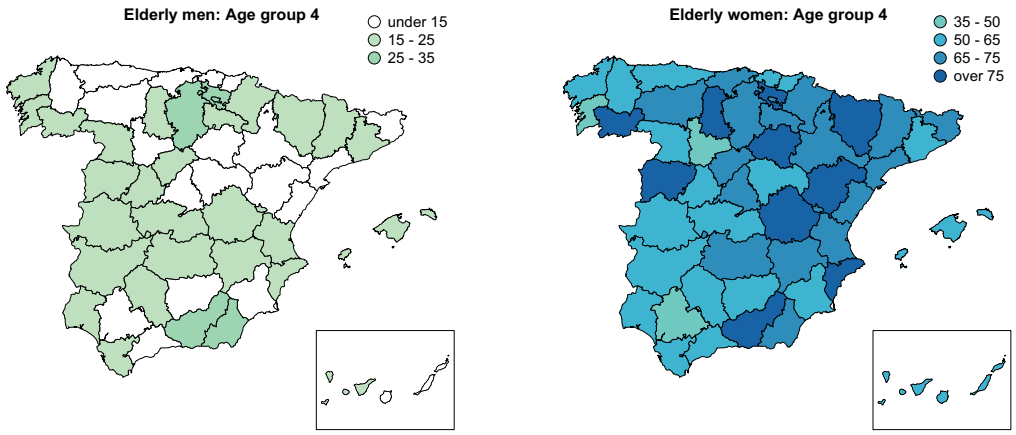


Figure 6.8. *Percentages of single-person households for elderly men (left) and women (right).*

7. Conclusions

Households are a key unit in a country's socioeconomic decision-making. Therefore, statistical studies of household composition in small and disaggregated areas are of great interest. Against this background, this paper addresses the prediction of total counts and proportions of single-person households by province, sex and age group of the main breadwinner. Given the difficulty of detecting single-person households from survey data, it is also important to model the disaggregated probabilities of false zeros. To do so, it has been taken into account that area-level zero-inflated PO mixed models are quite flexible to predict and explain count variables. In addition, they successfully model zero-inflated outcomes and have been applied in many fields of research. Consequently, the paper deals with an important, common but rather underestimated issue in SAE, which is the problem of zero inflation data.

To fit the model, we have calculated ML estimators of the model parameters and modal predictors of random effects by applying the ML-Laplace approximation. Then, we have considered the EBP, ESP and IN predictors. In theory, the EBP is very attractive because of its properties of approximately null bias and small RMSE. However, its formula contains double products of exponentials and integrals in $\mathbb{R}^{1+IJ}$. The evaluation

of exponential functions usually cause overflow problems when the observed counts are large, which is quite common under Option 1. This produces computational instability problems, especially when applying bootstrap resampling procedures, which made it necessary to omit the EBP from our simulation experiments. Finally, we have investigated the behaviour of the remaining predictors by generating the target variable from the same model as the one selected in the application to real data. Ultimately, we found that the ESP seems very attractive as it has a very low bias, but the IN predictor seems more interesting, as it has a small RMSE and lower computational cost. That is why we have decided to use the IN predictor in the case study. Regarding MSE estimation, we propose a parametric bootstrap procedure and recommend to use $B = 600$ iterations as a good tradeoff between accuracy and computational time.

Simulations also empirically investigated what happens if excess zeros are ignored in the prediction. Namely, if the excess of zeros is large, predictions based on the PO mixed model are rather inefficient. According to our results, the same applies if constant zero-inflated probabilities are considered, so that age-group randomness is required.

Section 6 presents an application to the 2016SHBS and illustrates how to use the proposed methodology. It has been concluded that living alone is a common residential choice across all age groups, influenced by marital separations, emancipation of children, cohabiting relationships and lifestyle in general. Declining fertility and increasing life expectancy are leading to an ageing population. Therefore an overwhelming increase in the proportion of single-person households is expected. What is more, differences in household composition for men and women are more pronounced among the elderly. In addition, RRMSE estimates are below 30% in most domains, which is a fairly good accuracy for a SAE problem.

Acknowledgements

Funding Body: This work has been supported by the Spanish Ministry of Universities, through the project PGC2018-096840-B-I00 and by the Valencian Government, through the project PROMETEO-2021-063. It has also benefited from a study grant from the Manuel Ventura Figueroa Foundation.

References

Anggreyani, A., Indahwati, I. and Kurnia, A. (2015). Small area estimation for estimating the number of infant mortality using a mixed effects zero inflated Poisson model. *Indonesian Journal of Statistics*, 20, 2, 108-115.

Berg, E.J. (2022). Empirical best prediction of small area means based on a unit-level Gamma-Poisson model. *Journal of Survey Statistics and Methodology*, 11, 4, 873-894.

Berg, E.J. and Fuller, W.A. (2012). Estimators of error covariance matrices for small area prediction. *Computational Statistics and Data Analysis*, 56, 10, 2949-2962.

Boubeta, M., Lombardía, M.J. and Morales, D. (2016). Empirical best prediction under area-level Poisson mixed models. *TEST*, 25, 548-569.

Boubeta, M., Lombardía, M.J. and Morales, D. (2017). Poisson mixed models for studying the poverty in small areas. *Computational Statistics and Data Analysis*, 107, 32-47.

Bugallo, M., Esteban, M.D., Marey-Pérez, M.F. and Morales, D. (2023). Wildfire prediction using zero-inflated negative binomial mixed models: Application to Spain. *Journal of Environmental Management*, 328, 116788.

Cai, S. and Rao, J.N.K. (2022). Selection of auxiliary variables for three-fold linking models in small area estimation: A simple and effective method. *Stats*, 5, 1, 128-138.

Chambers, R., Salvati, N. and Tzavidis, N. (2012). *M-quantile regression for binary data with application to small area estimation.* Centre for Statistical and Survey Methodology, University of Wollongong.

Chambers, R., Salvati, N. and Tzavidis, N. (2016). Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. *Journal of the Royal Statistical Society, Series A*, 179, 2, 453-479.

Chandra, H., Bathla, H.V.L. and Sud, U.C. (2010). Small area estimation under a mixture model. *Statistics in Transition*, 11, 3, 503-516.

Chandra, H. and Chambers, R. (2011). Small area estimation for skewed data in presence of zeros, *Calcutta Statistical Association Bulletin*, 63, 249-252.

Chandra, H. and Sud, U.C. (2012). Small area estimation for zero inflated data. *Communications in Statistics-Simulation and Computation*, 41, 632-643.

Chen, S. and Lahiri, P. (2012). Inferences on small area proportions. *Journal of the Indian Society of Agricultural Statistics*, 66, 121-124.

Cho, Y.K., Shim, K.W., Suk, H.W., Lee, H.S., Lee, S.W., Byun, A.R. and Lee, H.N. (2019). Differences between one-person and multi-person households on socioeconomic status, health behaviour, and metabolic syndrome across gender and age groups. *Korean Journal of Family Practice*, 9, 373-82.

Cohen, P.N. (2021). The rise of one-Person households. *Socius*, 7.

Datta, G.S. and Mandal, A. (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, 110, 512, 1735-1744.

Dreassi, E., Petrucci, A. and Rocco, E. (2014). Small area estimation for semicontinuous skewed spatial data: An application to the grape wine production in Tuscany. *Biometrical Journal*, 56, 1, 141-156.

Esteban, M.D., Morales, D., Pérez, A. and Santamaría, L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis*, 56, 2840-2855.

Fritsch, N.S., Riederer, B. and Seewann, L. (2023). Living alone in the city: Differentials in subjective well-being among single households 1995-2018. *Applied Research in Quality of Life*, 18, 2065-2087.

Ghosh, M., Kim, D., Sinha, K., Maiti, T., Katzoff, M. and Parsons, V.L. (2009). Hierarchical and empirical Bayes small domain estimation and proportion of persons without health insurance for minority subpopulations. *Survey Methodology*, 35, 53-66.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D. and Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics and Data Analysis*, 51, 2720-2733.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D. and Santamaría, L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Computational Statistics and Data Analysis*, 52, 5242-5252.

Greitemeyer, T. (2009). Stereotypes of singles: Are singles what we think? *European Journal of Social Psychology*, 39, 3, 368-383.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling".

Hartono, B., Kurnia, A. and Indahwati, I. (2017). Zero inflated binomial models in small area estimation with application to unemployment data in Indonesia. *International Journal of Computer Science and Network*, 6, 6, 746-752.

Hobza, T. and Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. *Journal of Official Statistics*, 32, 3, 661-69.

Hobza, T., Morales, D. and Santamaría, L. (2018). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *TEST*, 27, 2, 270-294.

Krieg, S., Boonstra, H.J. and Smeets, M. (2016). Small area estimation with zero-inflated data: A simulation study. *Journal of Official Statistics*, 32, 4, 963-986.

Lee, S.J. and Lee, S.H. (2021). Comparative analysis of health behaviours, health status, and medical needs among one-person and multi-person household groups: Focused on the ageing population of 60 or more. *Korean Journal of Family Medicine*, 42, 2, 73-83.

Lesthaeghe, R. (2014). The second demographic transition: A concise overview of its development. *Proc. of the National Academy of Sciences*. 111, 51, 18112-18115.

Liu, B. and Lahiri, P. (2017). Adaptive hierarchical Bayes estimation of small area proportions. *Calcutta Statistical Association Bulletin*, 69, 2, 150-164.

López-Vizcaíno, E., Lombardía, M.J. and Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, 13, 2, 153-178.

López-Vizcaíno, E., Lombardía, M.J. and Morales, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Association, Series A*, 178, 3, 535-565.

Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58, 308-325.

Marhuenda, Y., Morales, D. and Pardo, M.C. (2014). Information criteria for Fay-Herriot model selection. *Computational Statistics and Data Analysis*, 70, 268-280.

Michael, F. and Thomas, D. (2016). *Discrete data analysis with R: Visualization and modeling techniques for categorical and count data*. Chapman and Hall.

Militino, A.F., Ugarte, M.D. and Goicoa, T. (2015). Deriving small area estimates from information technology business surveys. *Journal of the Royal Statistical Society, Series A*, 178, 4, 1051-1067.

Molina, I., Saei, A. and Lombardía, M.J. (2007). Small area estimates of labour force participation under multinomial logit mixed model. *The Journal of the Royal Statistical Society, Series A*, 170, 975-1000.

Morales, D., Pagliarella, M.C. and Salvatore, R. (2015). Small area estimation of poverty indicators under partitioned area-level time models. *SORT*, 39, 1, 19-34.

Morales, D., Esteban, M.D., Pérez, A. and Hobza, T. (2021). *A course on small area estimation and mixed models*. Springer.

Morales, D., Krause, J. and Burgard, J.P. (2022). On the use of aggregate survey data for estimating regional major depressive disorder prevalence. *Psychometrika*, 87, 4.

Ortiz-Ospina, E. (2019). The rise of living alone: How one-person households are becoming increasingly common around the world. *Our World in Data*.

Park, B.Y., Kwon, H.J., Ha, M.N. and Burm, E.A. (2016). A comparative study on mental health between elderly living alone and elderly couples: Focus on gender and demographic characteristics. *Journal of Korean Public Health Nursing*, 20, 195-20.

Pfeffermann, D., Terryn, B. and Moura, F.A.S. (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Survey Methodology*, 34, 2, 235-249.

Sadik, K., Anisa, R. and Aqmaliyah, E. (2019). Small area estimation on zero-inflated data using frequentist and Bayesian approach. *Journal of Modern Applied Statistical Methods*, 18, 1, eP2677.

Snell, K.D.M. (2017). The rise of living alone and loneliness in history. *Social History*, 42, 1, 2-28.

Sugasawa, S., Kubokawa, T. and Ogasawara, K. (2017). Empirical uncertain bayes methods in area-level models. *Scandinavian Journal of Statistics*, 44, 3, 684-706.

Torabi, M. and Rao, J.N.K. (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis*, 127, 36-55.

Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E. and Chambers, R. (2015). Robust small area prediction for counts. *Statistical Methods in Medical Research*, 24, 3, 373-395.

Zhang, L.C. and Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society, Series B*, 66, 2, 479-496.

Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A. and Smith, G.M. (2009). *Mixed effects models and extensions in ecology with R*. Chapter 11. Springer.

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.

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.

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

Copyright notice and author opinions

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.