SORT 46 (2) July-December 2022, 217-244 DOI: 10.2436/20.8080.02.124

Missing data analysis and imputation via latent Gaussian Markov random fields

Virgilio Gómez-Rubio[1], Michela Cameletti[2] and Marta Blangiardo[3]

Abstract

This paper recasts the problem of missing values in the covariates of a regression model as a latent Gaussian Markov random field (GMRF) model in a fully Bayesian framework. The proposed approach is based on the definition of the covariate imputation sub-model as a latent effect with a GMRF structure. This formulation works for continuous covariates but for categorical covariates a typical multiple imputation approach is employed. Both techniques can be easily combined for the case in which continuous and categorical variables have missing values. The resulting Bayesian hierarchical model naturally fits within the integrated nested Laplace approximation (INLA) framework, which is used for model fitting. Hence, this work fills an important gap in the INLA methodology as it allows to treat models with missing values in the covariates. As in any other fully Bayesian framework, by relying on INLA for model fitting it is possible to formulate a joint model for the data, the imputed covariates and their missingness mechanism. In this way, it is possible to tackle the more general problem of assessing the missingness mechanism by conducting a sensitivity analysis on the different alternatives to model the non-observed covariates. Finally, the proposed approach is illustrated in two examples on modeling health risk factors and disease mapping.

MSC: *62D10, 62F15, 62M30.*

Keywords: *Imputation, missing values, GMRF, INLA, sensitivity analysis.*

[1] Department of Mathematics, School of Industrial Engineering, Albacete, Universidad de Castilla-La Mancha (Spain)

[2] Department of Economics, University of Bergamo, Bergamo, IT.

[3] Department of Epidemiology and Biostatistics, Imperial College London, London, UK.

Received: October 2022
Accepted: November 2022

1. Introduction

Missing data is an important issue a researcher needs to deal with in any statistical analysis; failing to properly account for it can result in a reduction of statistical power, or even in biased statistical inference. Consequently, countless methods have focused on how to deal with missing data (see, for example, **?**Enders, 2010; van Buuren, 2012; Trivellore, 2015; Little and Rubin, 2019).

Missing data can occur for a number of reasons, as described in Little and Rubin (2019). Sometimes, the missingness mechanism is ignorable and inference can rely on the observed data alone, appropriately coupled with a suitable imputation or data augmentation model if needed. When the missingness mechanism is not ignorable, a joint approach is required to fit the analysis model, impute the missing values and assess the missingness mechanism. Under this scenario, it is recommended that a sensitivity analysis is carried out to assess the impact of the missingness mechanism on the model parameters estimates (Mason et al., 2012).

The Bayesian paradigm has gained popularity for dealing with missing data, making no distinction between parameters and missing data which are considered as additional unknown parameters. For these reasons, and differently from other ad-hoc methods (Nakagawa, 2015), with a fully Bayesian approach it is possible to combine the analysis and imputation model in a joint estimation framework (Erler et al., 2016). For instance, Mason (2009) and Mason et al. (2012) developed a fully Bayesian missing imputation framework in order to adjust for several missing covariates in longitudinal or cross-sectional studies; each of the missing covariates is assigned an imputation model, all jointly modelled with the analysis model.

The approach we propose in this paper is based on recasting the imputation model to define it as a latent Gaussian Markov random field (GMRF, Rue and Held, 2005) which is part of a larger Bayesian hierarchical model. This fits naturally within the integrated nested Laplace approximation (INLA, Rue, Martino and Chopin, 2009) methodology, as an alternative to Markov chain Monte Carlo (MCMC, see, for example, Brooks et al., 2011). This approach is suitable for continuous covariates and can be also extended to impute categorical variables. This makes model fitting with missing covariates possible in INLA, and our new approach fills an important gap, as INLA has always required that the data in the latent GMRF defining the model to be fully observed. Here we focus on the case of missing values in the covariates as INLA can easily fit models with missing values in the response variable, simply computing the corresponding posterior predictive distribution derived from the analysis model to be fit (see, for example, Gómez-Rubio, 2020).

A previous attempt to solve the issue of missing values in the covariates in the INLA framework can be found in Gómez-Rubio and Rue (2018). They adopt a Gaussian prior for the imputation of the missing values in the covariates and sample from the missing data posterior distribution through INLA within MCMC. A different approach is proposed in Chapter 8 of Blangiardo and Cameletti (2015), where a bivariate model

for spatially misaligned data is estimated by adopting the stochastic partial differential equations (SPDE) approach Lindgren, Rue and Lindstrom (2011). Covariate values are imputed (in new locations) by assuming a spatial Gaussian field which is also included in the linear predictor of the response model. See also Barber et al. (2016); Forlani et al. (2020) for model examples on the use of spatial models for misalignment. Alternatively, Gómez-Rubio (2020) proposes a multiple imputation (MI) approach (Rubin, 1987, 1996; Carpenter and Kenward, 2012): the covariates are imputed multiple times through resampling, so that N complete datasets are used in the analysis model. All the results are then combined to obtain the final estimates of the model parameters (see Rubin, 1987, for details). The approach introduced here differs from previous approaches in that a joint framework is proposed, similarly to Mason et al. (2012). Through the joint model, the uncertainty about the imputation of the missing covariates propagates throughout the model so that it also reflects on the model parameters estimates in the analysis. At the same time, information from the outcome in the analysis model feedbacks on the imputation, making it unnecessary to include the outcome in the imputation model, as commonly done in the classic MI approach. This new approach fits naturally within the INLA framework, can be extended to consider different types of problems (i.e., not only spatial models) and can be easily fit with the associated `R-INLA` package for the `R` programming language (Gómez-Rubio, 2020).

The paper is structured as follows. In Section 2 we review methods for missing values, while in Section 3 we introduce our novel method for missing values imputation. Section 4 presents a brief summary of the INLA approach to Bayesian inference and how our novel approach fits within this framework. Section 5 shows two examples for the application of our proposed method and Section 6 presents discussion points.

2. Approaches to deal with missing data

In their seminal book, Little and Rubin (2019) identify three possible mechanisms of missingness. If the probability of being missing is the same for all the observations, we can assume that the missing data distribution does not depend on any of the observed or missing variables. In this case the data are said to be *missing completely at random* (MCAR). If the distribution of the missing data depends on completely observed variables (i.e., observed for all the subjects) and it does not depend on the varibles with missing values, the data are called *missing at random* (MAR). An example of MAR is that women are less likely to answer questions related to their income than men, but this has nothing to do with the income itself. Finally, if neither MCAR or MAR holds, the *missing not at random* (MNAR) case occurs and the missing values distribution depends on both missing and observed variables. For instance, in a neurological questionnaire, a subject is less likely to answer questions related to the disease if this is severe.

Under MCAR or MAR, the missing data mechanism is *ignorable*. As reported in Seaman and White (2013), this means that inferences obtained from a parametric model for the observed data alone are the same as inferences obtained from a joint model for

the data and missingness mechanism. On the contrary, if the data are MNAR the missing data mechanism is not ignorable and a model for the missingness mechanism is required. It is important to note that we cannot gather evidence from the data at hand about the missing data mechanism (MCAR, MNAR or MAR). On the basis of the knowledge regarding the data collection methods and the assumed relationship among the collected variables, it is possible only to make assumptions about the reasons for missing data, choose the best corresponding strategy for data analysis (Pigott, 2001) and conduct a sensitivity analysis on these assumptions (Mason et al., 2012).

The simplest and most popular ad-hoc method to deal with missing information consists in replacing the missing data with a plausible value, such as the mean or median calculated over the observed cases (or the mode if the variable is categorical) or to perform a complete cases analysis (i.e., removing the observations with one or more missing values). However, while the first method has the potential of distorting the data distribution and of underestimating their variability, the second one has the major drawback of reducing the power of the study (as the dataset for the analysis will have a reduced size) and of producing biased estimates if the MCAR assumption is not valid. To overcome this issue, inverse probability weighting was developed, based on the idea of assigning different weights to the different complete cases based on specific characteristics which are relevant for the missing data; in two reviews Carpenter, Kenward and Vansteelandt (2006) and Seaman and White (2013) showed advantages and drawbacks of the approach.

In the last three decades model-based methods have been preferred to account for missing data in the case of an ignorable missing data mechanism; see, for instance, the papers by Little 1992, Little and Rubin 2019 and Schafer and Graham 2002. Regression mean imputation is the simplest of the model-based methods, where the variable with missing data is predicted based on a regression model which includes the other variables as regressors. To overcome the issue of unreasonable lack of uncertainty for the imputed values, stochastic regression imputation was proposed to generate imputed values adding some random noise (Nakagawa, 2015).

A well established and increasingly popular model-based approach to dealing with missing data occurring in more than one variable is MI proposed by Rubin (1987, 1996). Through Monte Carlo simulation, it produces several versions of the complete dataset which only differ in the imputed missing values. Then, for each complete dataset the estimates of interest are computed by fitting the analysis model (also called *substantive model*) and the results are pooled together into a final estimate which takes into account the uncertainty of the imputed data. The imputation of the missing values can be done using mainly two strategies (van Buuren, 2012): i) *joint modeling*, when missing values are imputed by sampling from a multivariate model fitted to the data, for which usually a multivariate Gaussian is used (**?**Mason et al., 2012); ii) *fully conditional specification* (also known as multiple imputation using chained equations, MICE (van Buuren and Groothuis-Oudshoorn, 2011)), when conditional univariate distributions are used to impute the missing values iteratively through a variable-by-variable approach (see White, Royston and Wood 2011 for a thorough review of this method).

2.1. Bayesian inference

Bayesian inference provides a suitable framework for dealing with missing data, as it treats missing data similarly to model parameters, making no distinction between them. For these reasons and differently from other methods, with a fully Bayesian approach, it is possible to include the analysis model, imputation model and missingness mechanism model in a joint estimation framework (Erler et al., 2016).

Let $\mathcal{D}$ denote the *complete* set of data, which will include the response variable and the covariates. It is assumed that $\mathcal{D} = (\mathcal{D}_{obs}, \mathcal{D}_{mis})$, where $\mathcal{D}_{obs}$ denotes the observed values while $\mathcal{D}_{mis}$ refers to the missing values. Moreover, let $\boldsymbol{M}$ be the missing data indicator variable, i.e., a vector or matrix with the same length or dimension as $\mathcal{D}$ with values equal to 1 if the corresponding values of $\mathcal{D}$ is missing (and 0 otherwise).

Following the selection model approach (Nakagawa, 2015), the joint distribution of $\mathcal{D}$, $\boldsymbol{M}$, the model parameters $\boldsymbol{\theta}_{\mathcal{D}}$ and the parameters in the missingness model $\boldsymbol{\theta}_M$ can be expressed as

$$\begin{aligned}\pi(\mathcal{D}, \boldsymbol{M}, \boldsymbol{\theta}_{\mathcal{D}}, \boldsymbol{\theta}_M) &= \pi(\mathcal{D}, \boldsymbol{\theta}_{\mathcal{D}})\pi(\boldsymbol{M} \mid \mathcal{D}, \boldsymbol{\theta}_M)\pi(\theta_M) = \\ &= \pi(\boldsymbol{M} \mid \mathcal{D}, \boldsymbol{\theta}_M)\pi(\mathcal{D} \mid \boldsymbol{\theta}_{\mathcal{D}})\pi(\boldsymbol{\theta}_{\mathcal{D}})\pi(\theta_M).\end{aligned}$$

This formulation assumes that parameters $\boldsymbol{\theta}_{\mathcal{D}}$ and $\boldsymbol{\theta}_M$ are distinct and with independent priors and that the distribution of $\mathcal{D}$ (given $\boldsymbol{\theta}_{\mathcal{D}}$) does not depend on the parameters of the missingness model θ_M. Note that term $\pi(\boldsymbol{M} \mid \mathcal{D}, \boldsymbol{\theta}_M)$ represents the missingness model and $\pi(\mathcal{D} \mid \boldsymbol{\theta}_{\mathcal{D}})$ the likelihood of the data.

Following this, $\pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \mathcal{D}_{mis}, \boldsymbol{\theta}_M)$ depends on a set of parameters $\boldsymbol{\theta}_M$, and models the missing data mechanism for the three cases introduced above (Little and Rubin, 2019):

MCAR, if the distribution does not depend on any of the fully or partially observed variables, i.e., $\pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \mathcal{D}_{mis}, \boldsymbol{\theta}_M) = \pi(\boldsymbol{M} \mid \boldsymbol{\theta}_M)$.

MAR, if the distribution depends only on fully observed variables, which means that $\pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \mathcal{D}_{mis}, \boldsymbol{\theta}_M) = \pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \boldsymbol{\theta}_M)$. This implies that, given the observed data, the missingness mechanism does not depend on the unobserved data.

MNAR, if the distribution $\pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \mathcal{D}_{mis}, \boldsymbol{\theta}_M)$ depends on fully and partially observed variables.

If the data are MCAR or MAR and the parameters $\boldsymbol{\theta}_M$ are distinct of the parameters of the data generating process, $\boldsymbol{\theta}_{\mathcal{D}}$, and with independent priors, then the missing data mechanism is *ignorable* and $\pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \mathcal{D}_{mis}, \boldsymbol{\theta}_M)$ can be omitted (Seaman and White, 2013). On the contrary if the data are MNAR, the missing data mechanism is not ignorable and a model for missingness is required (i.e., a logistic model) and has to be jointly estimated with the main model, that will include an imputation model for the missing values.

Note that it is not possible to tell from the data at hand whether the missing observations are MCAR, MNAR or MAR and at the same time it is not trivial to specify a model of missingness. In this case, a sensitivity analysis needs to be carried out to assess the impact of different scenarios for the missing data on the estimates of the model parameters (Carpenter, Kenward and White, 2007; Mason et al., 2012).

2.2. Missing data in the response variable

Let now $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{x})$ be the set of data including the response $\boldsymbol{y}$ and the covariates $\boldsymbol{x}$. If it is assumed that the covariates are fully observed and that the response variable $\boldsymbol{y}$ contains missing values, i.e., the response variable $\boldsymbol{y}$ can be split into observed values, $\boldsymbol{y}_{obs}$, and unobserved values, $\boldsymbol{y}_{mis}$. Hence, $\mathcal{D}_{obs} = (\boldsymbol{y}_{obs}, \boldsymbol{x})$ and $\mathcal{D}_{mis} = (\boldsymbol{y}_{mis})$. In this case likelihood $\boldsymbol{\pi}(\mathcal{D}_{obs}, \mathcal{D}_{mis} \mid \boldsymbol{\theta}_{\mathcal{D}})$ corresponds to the distribution of $\boldsymbol{\pi}(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis} \mid \boldsymbol{x}, \boldsymbol{\theta}_y)$, with $\boldsymbol{\theta}_y$ the hyperparameters in the likelihood.

If we assume that the missing data mechanism is ignorable, the imputation of the missing data values $\boldsymbol{y}_{mis}$ is simply done through the posterior predictive distribution $\boldsymbol{\pi}(\boldsymbol{y}_{mis} \mid \boldsymbol{y}_{obs}, \boldsymbol{x})$. In general, we will have the observation model by defining an appropriate distribution for the likelihood. In addition, the mean of observation i, ϕ_i ,will be linked to a linear predictor η_i on the covariates and other effects using an appropriate link function $g(\cdot)$, i.e.,

$$g(\phi_i) = \eta_i = \beta_0 + \sum_{p=1}^{P} \beta_p x_{pi} + \sum_{l=1}^{L} f_l(u_{li}). \tag{1}$$

Here, β_0 is an intercept, $\{\beta_p\}_{p=1}^{P}$ the coefficients of the P covariates available $\{\boldsymbol{x}_p\}_{p=1}^{P}$ and $\{f_l(\cdot)\}_{l=1}^{L}$ represents L different non-linear effects on covariates $\{\boldsymbol{u}_l\}_{l=1}^{L}$ (which are also part of the observed data $\mathcal{D}_{obs}$ now).

If the data are MNAR, a missing mechanism model $\boldsymbol{\pi}(\boldsymbol{M} \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}_M)$ is required in addition to the previous model, e.g.,

$$\begin{aligned} M_i \mid p_i &\sim Bernoulli(p_i) \\ logit(p_i) &= \gamma_0 + \sum_{r=1}^{R} \gamma_r x_{ri} + \delta y_i \end{aligned} \tag{2}$$

where $\boldsymbol{\theta}_M = [\gamma_1, \gamma_1 \cdots \gamma_R \ \delta]^{\intercal}$ and M_i is a missingness indicator for y_i. In addition, an imputation model for the missing values will be required. Furthermore, δ is a coefficient that measures the effect of the response variable on the missingness mechanism.

However, in this work we will assume that there are no missing observations in the response or that the missingness mechanism is ignorable, which means that posterior inference is based on the predictive distribution.

2.3. Missing data in the covariates

We now consider the case when $\mathcal{D}_{obs} = (\boldsymbol{y}, \boldsymbol{x}_{obs})$ and $\mathcal{D}_{mis} = (\boldsymbol{x}_{mis})$, with $\boldsymbol{x}_{obs}$ the observed values of the covariates and $\boldsymbol{x}_{mis}$ the missing ones. Henceforth, the likelihood

$\pi(\mathcal{D}_{obs}, \mathcal{D}_{mis} \mid \boldsymbol{\theta}_{\mathcal{D}})$ can be written as

$$\pi(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{x}_{mis} \mid \boldsymbol{\theta}_{\mathcal{D}}) = \pi(\boldsymbol{y} \mid \boldsymbol{x}_{obs}, \boldsymbol{x}_{mis}, \boldsymbol{\theta}_y)\pi(\boldsymbol{x}_{obs}, \boldsymbol{x}_{mis} \mid \boldsymbol{\theta}_x)$$

assuming that $\boldsymbol{\theta}_{\mathcal{D}} = [\boldsymbol{\theta}_y^{\mathsf{T}}\ \boldsymbol{\theta}_x^{\mathsf{T}}]^{\mathsf{T}}$ is the vector of conditionally independent parameters. The distribution $\pi(\boldsymbol{x}_{obs}, \boldsymbol{x}_{mis} \mid \boldsymbol{\theta}_x)$ represents the joint distribution of observed and missing covariates and it includes the imputation model. For example, the joint distribution can be a multivariate normal distribution (taking into consideration correlation between covariates) for continuous covariates, or a discrete distribution if the covariate is categorical.

In general, we will have the observation model with a linear predictor as in Equation (1) together with the imputation model and the missingness model (described in Section 3) as in Equation (2) but only if the missing data are MNAR.

3. Imputation of continuous missing covariates

Differently from Section 2, let $\boldsymbol{z} = [\boldsymbol{z}_{obs}{}^{\mathsf{T}}\ \boldsymbol{z}_{mis}{}^{\mathsf{T}}]^{\mathsf{T}}$ denote now the complete set of values of a covariate, which will typically be a column vector. The response values $\boldsymbol{y}$ will be written separately where needed. This is done for simplicity, so that the imputation of a single covariate with missing observations will be considered now. However, this approach can be easily extended to consider the imputation of missing values in several continuous covariates using a multivariate model.

Let $\boldsymbol{z}^*$ be a latent effect that is split in two parts, i.e., $\boldsymbol{z}^* = [\boldsymbol{z}^*_{obs}{}^{\mathsf{T}}\ \boldsymbol{z}^*_{mis}{}^{\mathsf{T}}]^{\mathsf{T}}$. The main idea is to define latent effect $\boldsymbol{z}^*$ as a GMRF with mean $\boldsymbol{\mu}^*(\boldsymbol{\theta}_I)$ and precision $\boldsymbol{Q}^*(\boldsymbol{\theta}_I)$ so that $\boldsymbol{z}^*_{obs}$ is as close as possible to the actual values $\boldsymbol{z}_{obs}$ and so that $\boldsymbol{z}^*_{mis}$ is obtained using a particular imputation model for $\boldsymbol{z}_{mis}$ that depends on observed covariates $\boldsymbol{z}_{obs}$ and some parameters $\boldsymbol{\theta}_I$.

To guarantee that the distribution of $\boldsymbol{z}^*_{obs}$ is taken to be as close as possible to the observed covariate data $\boldsymbol{z}_{obs}$, the mean of $\boldsymbol{z}^*_{obs}$ is set equal to $\boldsymbol{z}_{obs}$ and its associated sub-block in $\boldsymbol{Q}^*(\boldsymbol{\theta}_I)$ equal to a diagonal matrix with high values (e.g., 10^{10}) in the diagonal. In this way, the values of $\boldsymbol{z}^*_{obs}$ are centered at observed values $\boldsymbol{z}_{obs}$ and have a negligible variation about these observed values. Regarding the distribution of $\boldsymbol{z}^*_{mis}$ (with mean $\boldsymbol{\mu}_c$ and precision $\boldsymbol{Q}_c$), it will be based on an imputation model on observed covariates $\boldsymbol{z}_{obs}$ and parameters $\boldsymbol{\theta}_I$. Finally, we will also assume that $\boldsymbol{z}^*_{obs}$ and $\boldsymbol{z}^*_{mis}$ are independent because the marginal distribution of $\boldsymbol{z}^*_{mis}$ will include all dependence of the missing values on the observed data $\boldsymbol{z}_{obs}$.

Consequently, the joint distribution of $\boldsymbol{z}^*$ is given by

$$\boldsymbol{z}^* \mid \boldsymbol{\theta}_I \sim \text{Normal}\left(\begin{bmatrix} \boldsymbol{z}_{obs} \\ \boldsymbol{\mu}_c \end{bmatrix}, \begin{bmatrix} 10^{10}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & Q_c \end{bmatrix}\right), \tag{3}$$

where $\boldsymbol{I}$ represents the identity matrix. The distribution of $\boldsymbol{z}^*$ will be used later when defining the imputation model for the missing values as a latent effect for `R-INLA` in Section 4.

3.1. Imputation latent effect

To derive the distribution for the imputation model, $\pi(z_{mis} \mid z_{obs}, \theta_I)$, a multivariate Normal distribution is assumed for the joint distribution of the complete set of covariates z:

$$z \mid \theta_I \sim \text{Normal}\left(\begin{bmatrix} \mu_{obs} \\ \mu_{mis} \end{bmatrix}, \begin{bmatrix} Q_{obs,obs} & Q_{obs,mis} \\ Q_{mis,obs} & Q_{mis,mis} \end{bmatrix}\right) = \text{Normal}(\mu, Q), \quad (4)$$

where both the mean and the precision matrix can depend on θ_I. It follows that the *imputation model* is defined by the following conditional distribution (Rue and Held, 2005):

$$z_{mis} \mid z_{obs}, \theta_I \sim \text{Normal}(\mu_c, Q_c)$$

where $\mu_c = \mu_{mis} - Q_{mis,mis}^{-1} Q_{mis,obs}(z_{obs} - \mu_{obs})$ and $Q_c = Q_{mis,mis}$. Note that μ_c and Q_c are necessary to define the distribution of the new latent effect given in Equation (3).

As stated above, the distribution of z_{mis}^* will play the role of the imputation model of the missing values. This imputation model will, in practice, be a sub-model in a larger model that will be defined using the conditional distribution of the missing values z_{mis} given the observed data z_{obs} and hyperparameters θ_I. Note that in this sub-model z_{obs} can be regarded as the data while z_{mis} and θ_I are the parameters to estimate. Because this sub-model will be included as part of a fully Bayesian larger model, posterior inference on z_{mis} and θ_I will be based on all observed data in the model (i.e., response variable and observed covariates) so that there is feedback from other parts of the model to make inference on z_{mis} and θ_I.

Considering only the data and parameters in the sub-model, the way in which the imputation sub-model is defined relies on the distribution of z_{mis} given z_{obs}. This can be written as

$$\pi(z_{mis} \mid z_{obs}) = \int_{\Theta_I} \pi(z_{mis}, \theta_I \mid z_{obs}) d\theta_I = \int_{\Theta_I} \pi(z_{mis} \mid z_{obs}, \theta_I) \pi(\theta_I \mid z_{obs}) d\theta_I.$$

where Θ_I is the parametric space of θ_I.

Here, $\pi(z_{mis} \mid z_{obs}, \theta_I)$ is the conditional distribution of the missing values given the observed data and the hyperparameters of the imputation model introduced above. Also, note that $\pi(\theta_I \mid z_{obs})$ can be regarded as the distribution of the hyperparameters in the imputation sub-model given the observed data. Note that this distribution is estimated only from the observed data z_{obs}, so it can be regarded as an *informative prior* for θ_I. Moreover, it can be rewritten as

$$\pi(\theta_I \mid z_{obs}) \propto \pi(z_{obs} \mid \theta_I) \pi(\theta_I)$$

where $\pi(z_{obs} \mid \theta_I)$ is obtained by integrating z_{mis} out in the distribution of z, that is, $\pi(z_{obs} \mid \theta_I) = \int \pi(z_{obs}, z_{mis} \mid \theta_I) dz_{mis}$. Finally, the hyperparameters θ_I are typically modelled as exchangeable a priori.

Next, two particular examples of imputation with a typical linear regression and a spatial model (useful when the covariate is spatially correlated) are described. It is worth noting that the principles presented below can be extended to a wide range of models, including longitudinal data, time series and other smooth terms.

3.2. Imputation with a linear regression model

The first imputation model that we describe is based on the linear regression model. We assume that the mean of the multivariate Normal distribution in Equation (4) is defined, considering the n observations, as $\boldsymbol{X}\boldsymbol{\beta}$. Here, $\boldsymbol{X}$ is a matrix of P fully observed covariates (columnwise) with associated coefficient vector $\boldsymbol{\beta} = [\beta_0 \cdots \beta_P]^\top$. To match the structure of $\boldsymbol{z} = [\boldsymbol{z}_{obs}{}^\top \boldsymbol{z}_{mis}{}^\top]^\top$, matrix $\boldsymbol{X}$ can be rewritten as a block matrix as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{obs} \\ \boldsymbol{X}_{mis} \end{bmatrix}$$

Under the linear regression model, we assume that the mean of $\boldsymbol{z}$ depends on a linear combination of the fully observed covariates, i.e., $\boldsymbol{\mu} = E(\boldsymbol{z}) = \boldsymbol{X}\boldsymbol{\beta}$. By adopting the block notation, we thus assume that the joint distribution of Equation (4) is given by

$$\boldsymbol{z} \mid \boldsymbol{\theta}_I \sim \text{Normal}\left(\begin{bmatrix} \boldsymbol{X}_{obs}\boldsymbol{\beta} \\ \boldsymbol{X}_{mis}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \tau\boldsymbol{I}_{obs} & \boldsymbol{0} \\ \boldsymbol{0} & \tau\boldsymbol{I}_{mis} \end{bmatrix} \right),$$

where τ is the precision hyperparameter and $\boldsymbol{I}_{obs}$ and $\boldsymbol{I}_{mis}$ are identity matrices whose dimensions depend on the number of missing and observed data in $\boldsymbol{z}$. In this case the vector of hyperparameters is given by $\boldsymbol{\theta}_I = [\boldsymbol{\beta}^\top \tau]^\top$. Note that, given $\boldsymbol{\theta}_I$, observations are assumed independent of each other, which simplifies the model.

Following the approach presented in Section 3.1, we obtain that the conditional distribution of $\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\theta}_I$ (i.e., the imputation model) has the following mean and precision:

$$\boldsymbol{\mu}_c = \boldsymbol{X}_{mis}\boldsymbol{\beta}, \qquad \boldsymbol{Q}_c = \tau\boldsymbol{I}_{mis},$$

As stated above, note that $\boldsymbol{\beta}$ and τ are informed by $\pi(\boldsymbol{\beta}, \tau \mid \boldsymbol{z}_{obs})$, which is proportional to $\pi(\boldsymbol{z}_{obs} \mid \boldsymbol{\beta}, \tau)\pi(\boldsymbol{\beta}, \tau)$. Note that $\pi(\boldsymbol{z}_{obs} \mid \boldsymbol{\beta}, \tau)$ can be easily derived from the multivariate normal distribution of $\boldsymbol{z}$ above and that it will also be a multivariate normal distribution with mean $X_{obs}\boldsymbol{\beta}$ and precision $\tau\boldsymbol{I}_{obs}$.

Finally, priors must be set on the hyperparameters. For simplicity, each of the elements in $\boldsymbol{\beta}$ is assigned a Normal distribution with zero mean and small precision. Parameter τ has a vague prior (e.g., a Gamma distribution with small precision). All hyperparameters are independent a priori, so that $\pi(\boldsymbol{\theta}_I) = \pi(\tau)\Pi_{i=0}^{P}\pi(\beta_i)$. Note that other priors could be easily considered here.

3.3. Imputation with a spatial model

When the covariate to be imputed is spatially correlated we can assume a conditional autoregressive (CAR) specification (Held and Rue, 2010) so that the mean is $\boldsymbol{\mu} = \boldsymbol{\alpha} = [\alpha \cdots \alpha]^\top$ and the precision is $\boldsymbol{Q} = \tau(\boldsymbol{I} - \rho\boldsymbol{W})$. Here, α is the intercept of the linear predictor, ρ is a spatial autocorrelation parameter, and $\boldsymbol{W}$ is an adjacency matrix, defining the sets of neighbours. This is often scaled dividing it by its largest eigenvalue as this will allow us to take ρ in the $(0, 1)$ interval. Note that $\boldsymbol{W}$ can be rewritten as a block

matrix with four sub-matrices according to missing and observed values, as done with $\boldsymbol{Q}$ in Equation (4). The vector of hyperparameters is now given by $\boldsymbol{\theta}_I = [\tau \ \ \rho \ \ \alpha^\intercal]^\intercal$.

Adopting block notation, under the CAR specification for imputation the following joint distribution is assumed for $\boldsymbol{z} = [\boldsymbol{z}_{obs}{}^\intercal \ \boldsymbol{z}_{mis}{}^\intercal]^\intercal$:

$$\boldsymbol{z} \mid \boldsymbol{\theta}_I \sim \text{Normal}\left(\begin{bmatrix} \boldsymbol{\alpha}_{obs} \\ \boldsymbol{\alpha}_{mis} \end{bmatrix}, \begin{bmatrix} \tau(\boldsymbol{I}_{obs} - \rho \boldsymbol{W}_{obs,obs}) & -\tau\rho \boldsymbol{W}_{obs,mis} \\ -\tau\rho \boldsymbol{W}_{mis,obs} & \tau(\boldsymbol{I}_{mis} - \rho \boldsymbol{W}_{mis,mis}) \end{bmatrix} \right).$$

It then follows that the conditional distribution of $\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\theta}_I$ (i.e., the imputation model) is characterised by the following mean and precision matrix:

$$\boldsymbol{\mu}_c = \boldsymbol{\alpha}_{mis} - (\boldsymbol{I}_{mis} - \rho \boldsymbol{W}_{mis,mis})^{-1}(-\rho \boldsymbol{W}_{mis,obs})(\boldsymbol{z}_{obs} - \boldsymbol{\alpha}_{obs})$$

$$\boldsymbol{Q}_c = \tau\,(\boldsymbol{I}_{mis} - \rho \boldsymbol{W}_{mis,mis})$$

Again, τ, ρ and α are informed by $\pi(\tau, \rho, \alpha \mid \boldsymbol{z}_{obs})$, which is proportional to the product $\pi(\boldsymbol{z}_{obs} \mid \tau, \rho, \alpha)\pi(\tau, \rho, \alpha)$. As in the previous case, $\pi(\boldsymbol{z}_{obs} \mid \tau, \rho, \alpha)$ can be easily derived from the multivariate normal distribution of $\boldsymbol{z}$ above and that it will also be a multivariate normal distribution with mean α_{obs} and precision $\tau(\boldsymbol{I}_{obs} - \rho \boldsymbol{W}_{obs,obs})$.

Finally, α is given a Gaussian prior with zero mean and small precision, τ is assigned a vague prior (e.g., a Gamma distribution with a small precision), while $\text{logit}(\rho)$ is assigned a Gaussian prior with zero mean and small precision (see, for example, Gómez-Rubio, 2020, Chapter 5, for details on why this parameterisation is used).

3.4. Extension to the imputation of categorical missing covariates

The imputation of the missing values in categorical variables does not fit into the GMRF framework described in Section 3 as these variables are defined in a discrete space. For this reason, a different approach will be considered for defining the imputation model $\pi(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\theta}_I)$ and for estimating the model. In particular, as imputation model we will consider a multinomial likelihood which can be fit with INLA by using the multinomial-Poisson transformation (Baker, 1994).

Note that in this case the procedure is similar to the multiple imputation approach: the imputation model is specified where the categorical variables with missing values are considered as the response variables, so that the predictive distribution of the missing observations can be computed. Similarly to the case of missing data in the response, values are sampled to fill the missing values in the covariates. Then, the analysis model is run by using the imputed covariates as completely known. This procedure is repeated by simulating several samples and estimating the corresponding models; finally, all the resulting models are pooled by using Bayesian model averaging (Gómez-Rubio and Rue, 2018). Note that this approach does not produce feedback in the estimation of the parameters of the imputation model as in the joint approach, given that it is done in two-stages rather than jointly. For this reason, and similarly to the classical MI, the outcome $\boldsymbol{y}$ should be included in the imputation model. Alternatively, INLA within MCMC can be used to fit

the joint model using a fully Bayesian approach (see the example in Gómez-Rubio and Rue, 2018).

Inference on the model parameters when multiple imputation of a categorical covariate can be summarised as follows. Considering the generic parameter θ_k we can write its posterior marginal distribution as:

$$\pi(\theta_k \mid \boldsymbol{z}_{obs}, \boldsymbol{y}) = \sum_{\boldsymbol{z}_{mis} \in \Theta_{mis}} \pi(\theta_k, \boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{y}) = \sum_{\boldsymbol{z}_{mis} \in \Theta_{mis}} \pi(\theta_k \mid \boldsymbol{z}_{obs}, \boldsymbol{z}_{mis}, \boldsymbol{y}) \pi(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{y}).$$

Here, Θ_{mis} represents the parametric space of the missing values of the categorical covariate, which in a Bayesian framework are considered to be random variables.

Given L samples $\{\boldsymbol{z}_{mis}^{(l)}\}_{l=1}^{L}$ from $\pi(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{y})$, the previous marginal can be approximated as

$$\pi(\theta_k \mid \boldsymbol{z}_{obs}, \boldsymbol{y}) \simeq \frac{1}{L} \sum_{l=1}^{L} \pi(\theta_k \mid \boldsymbol{z}_{obs}, \boldsymbol{z}_{mis}^{(l)}, \boldsymbol{y}),$$

where $\pi(\theta_k \mid \boldsymbol{z}_{obs}, \boldsymbol{z}_{mis}^{(l)}, \boldsymbol{y})$ is the marginal of θ_k obtained from fitting the original model with the observed data and the imputed covariate $\boldsymbol{z}_{mis}^{(l)}$.

Note that when continuous covariates with missing values are also present both approaches can be combined. For example, an imputation model can be combined for the continuous covariate which is part of the joint model that is fit to every simulated dataset where only the missing values of the categorical covariate are filled in. Furthermore, a missingness model for the categorical variables can be incorporated into the model similarly to the one used for the continuous variables.

4. The Integrated Nested Laplace Approximation approach (INLA)

The approach presented in the previous sections can be implemented using a number of methods for Bayesian inference. However, it overcomes a major limitation in the INLA method as, at present, it cannot cope with missing values in covariates. An introduction to the INLA method and the computational details is presented here; then we focus on how to implement our proposed framework.

INLA (Rue et al., 2017; Martino and Riebler, 2019; Gómez-Rubio, 2020) is a deterministic approach for Bayesian inference. It is designed for the class of latent Gaussian Markov random field models, where the distribution of the response y_i (observed for the i-th unit) is assumed to belong to a distribution family (usually part of the exponential family). This is often characterized by a parameter ϕ_i (i.e., the mean of y_i) defined as a function of a structured additive predictor η_i through a link function such that $g(\phi_i) = \eta_i$ (e.g. the logarithm function is used for Poisson data). The linear predictor is defined as in equation (1).

Regarding the tems in the linear predictor, recall that β_0 is the intercept, coefficients $\boldsymbol{\beta} = [\beta_1 \cdots \beta_P]^\intercal$ quantify the (linear) effect of some covariates $\boldsymbol{x} = \{\boldsymbol{x}_p\}_{p=1}^{P}$ on the re-

sponse, and $\boldsymbol{f} = \left\{f^{(1)}(\cdot), \ldots, f^{(L)}(\cdot)\right.$ is a set of functions defined in terms of some covariates $\boldsymbol{u} = \{\boldsymbol{u}_l\}_{l=1}^{L}$.

Through functions $f(\cdot)$ it is possible to include in the model random effects (perhaps indexed in space and time), smooth and non-linear effects of the covariates. For this reason, the class of latent GMRF models can accommodate a wide range of models, from standard generalized linear models (GLM) to generalized linear mixed models (GLMM), including data for time series, lattice data, point pattern and geostatistical data.

As stated, the set of latent effects $\boldsymbol{\chi} = \{\boldsymbol{\eta}, \beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$ is a latent GMRF in the model, which depends on some hyperparameters $\boldsymbol{\theta}_2$. Moreover, observations are assumed to be independent given the latent effects $\boldsymbol{\chi}$ and the likelihood hyperparameters denoted by $\boldsymbol{\theta}_1$. For convenience, in the following the vector of hyperparameters wil be denoted as $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^{\mathsf{T}}\ \boldsymbol{\theta}_2^{\mathsf{T}}]^{\mathsf{T}}$.

The outputs of Bayesian inference with INLA are the marginal posterior distributions for each element of the latent effects and hyperparameters vector denoted by $p(\boldsymbol{\chi}_\bullet \mid \boldsymbol{y})$ and $p(\boldsymbol{\theta}_\bullet \mid \boldsymbol{y})$, respectively. INLA provides deterministically accurate approximations to these distributions in a short computing time by using the Laplace approximation and numerical integration.

Each latent GMRF model can be rewritten hierarchically with three levels:

1. The model for the observed data $\boldsymbol{y} = [y_1 \cdots y_n]^{\mathsf{T}}$ (i.e., the likelihood) defined as a function of some parameters $\boldsymbol{\chi}$ and hyperparameters $\boldsymbol{\theta}$:

$$\boldsymbol{y} \mid \boldsymbol{\chi}, \boldsymbol{\theta} \sim \pi(\boldsymbol{y} \mid \boldsymbol{\chi}, \boldsymbol{\theta}) = \prod_{i \in \{1,\ldots,n\}} \pi(y_i \mid \chi_i, \boldsymbol{\theta}).$$

2. The model for the latent effects $\boldsymbol{\chi}$:

$$\boldsymbol{\chi} \mid \boldsymbol{\theta} \sim \text{Normal}\,(\boldsymbol{0}, \boldsymbol{Q}(\boldsymbol{\theta}))$$

   where $\boldsymbol{Q}(\boldsymbol{\theta})$ is a sparse precision matrix given the GMRF assumption.

3. The model for the complete vector of hyperparameters: $\pi(\boldsymbol{\theta})$. As usually hyperparameters are assumed to be independent a priori, $\pi(\boldsymbol{\theta})$ will be defined as the product of different univariate prior distributions.

Given all these models and components the joint posterior distribution of the random effects and the hyperparameters is given by

$$\pi(\boldsymbol{\chi}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{y} \mid \boldsymbol{\chi}, \boldsymbol{\theta}) \pi(\boldsymbol{\chi} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}).$$

As stated above, INLA computes the posterior marginals of the hyperparameters and latent effects using that representation by means of numerical integration and the Laplace approximation (see Rue et al., 2009, for details).

4.1. Computational details

The INLA approach is implemented through an `R` package named `R-INLA`, which is available from the INLA website (http://www.r-inla.org/home). The model to be fit is defined by setting a formula with all the additive latent effects in the model, which includes fixed and random effects. The `R-INLA` package includes a good number of implemented latent effects but others can be implemented as well (see, for example Gómez-Rubio, 2020). Note that by default, when `R-INLA` finds missing values in the covariates (which have the value `NA` in `R`) they are replaced by zeros so that the effect of the covariate does not affect the linear prediction of that subject. However, this is an issue that could result in biased estimates of the coefficients of the covariates. This is described in the `R-INLA` list of frequently asked questions (FAQ) in the package website. If the missing value is found in the response variable, the predictive distribution is computed.

Generic latent effects can be implemented by defining their structure as a latent GMRF. This means defining the mean, precision, hyperparameters and the priors of the hyperparameters. These are known as `rgeneric` latent effects in `R-INLA` (see, for example Gómez-Rubio, 2020, Chapter 11). Once a new latent effect is defined, it can be easily incorporated as any other additive effect in the model formula.

For the new latent effects described in this paper and defined in Equation (3) we have to specify the mean $\boldsymbol{\mu}_c$ and precision $\boldsymbol{Q}_c$ of the block of the missing values. Remember that the block of the observed covariates is simply there to make those values of the latent effect to be as close as possible to the observed values and that it does not depend on any hyperparameter or other data. Furthermore, the role of the prior on the hyperparameters of the imputation model $\boldsymbol{\theta}_I$ is now taken by distribution $\pi(\boldsymbol{\theta}_I \mid \boldsymbol{z}_{obs})$. Hence, the actual prior used in the latent effects is taken as

$$\pi(\boldsymbol{\theta}_I \mid \boldsymbol{z}_{obs}) \propto \pi(\boldsymbol{z}_{obs} \mid \boldsymbol{\theta}_I)\pi(\boldsymbol{\theta}_I)$$

and the normalizing constant is ignored as it is not needed. In a standard implementation of a latent effect, the prior of $\boldsymbol{\theta}_I$ would be a typical distribution density that depends on a set of fixed hyperparameters, but now the prior of $\boldsymbol{\theta}_I$ is made of the product of the two terms above. For this reason, it can be regarded as an informative prior as it is essentially estimated from a model fit to $\boldsymbol{z}_{obs}$. This is what will allow the latent effect to produce good estimates of the missing values (if the imputation model is correct). In general, there is no way to assess this, but the more covariates used in the imputation model the better (see Gelman and Hill, 2007, Chapter 25). The actual prior of the model hyperparameters is $\pi(\boldsymbol{\theta}_I)$ and this can take different forms depending on the number and type of hyperparameters in the model. Usually, this will be split into the product of several univariate prior distributions.

Note also that `R-INLA` works with unbounded hyperparameters, so that the parameters in $\boldsymbol{\theta}_I$ may need to be transformed when the latent effect is defined. This may also require to include additional terms in the prior (see, for example Gómez-Rubio, 2020, Chapter 11). A typical example is to use internally the log-precision instead of the precision.

Once the imputation latent effect is included in the model formula, it will be part of the joint latent effect χ and incorporated into the Bayesian model, so that a full Bayesian approach is used to estimate all the model parameters.

As stated in previous sections, a missingness model can be included (in addition to an imputation one) for the case in which missingness is MAR or MNAR. Including a missingness model requires defining a model with two likelihoods: one for the main model and a binomial model for the missingness indicator variables. Note that under MCAR and MAR both models are independent, hence the latter is not needed; however, under MNAR it is necessary to explicitly include it and to make it dependent on the variables with imputed values. Hence, there will be feedback between both models that may affect the imputation process and the estimation of the other model parameters.

Full details about how to fit these models in `R` are provided in the associated R code (see Section 5 below). A new `MIINLA` R package which implements the approach proposed here and that can be easily used together with the `R-INLA` package is available at https://github.com/becarioprecario/MIINLA.

5. Examples

In this section we develop two examples to show how the imputation method proposed above can be used with INLA under MCAR, MAR and MNAR. The first example shows a typical regression model in biostatistics with real missing data. This is useful to show how a typical multiple linear regression can be used for multiple imputation. The second one is based on spatially correlated data to assess the performance of our proposal on a simulated study in which a spatially correlated covariate is missing. Note that the aim is not to provide a comprehensive analysis of the dataset with missing values but to illustrate the methods described in this paper.

All models have been fit with INLA and its associated `R` package `R-INLA`. The `R` code to reproduce the examples described here is available from a GitHub repository at https://github.com/becarioprecario/MIINLA_paper.

5.1. Imputation using linear models

The `nhanes2` dataset (**?**) in the `mice` `R` package (van Buuren and Groothuis-Oudshoorn, 2011) records data on 25 participants in the National Health and Nutrition Examination Survey (NHANES). Variables in the dataset include body mass index, cholesterol level, age group and hypertensive status. The dataset presents missing observations in body mass index, hypertensive status and cholesterol level.

We will use this dataset to build a model to explain cholesterol level on age group and body mass index, where this is imputed. The imputation model will be based on a linear regression on the age group. There are three age groups 20-39, 40-59 and 60+ years, and the first group will be set as the reference level.

It is worth noting that having missing values in the response variable (i.e., cholesterol level) is not a problem as the predictive distribution can be easily computed with INLA.

Hence, the output from fitting this model will include the posterior distribution of the imputed values as well as the predictive distribution for the missing responses.

The analysis model is the following:

$$chol_i = \beta_0 + \beta_1 age_i^{40-59} + \beta_2 age_i^{60+} + \beta_3 bmi_i + \varepsilon_i,\ i = 1, \ldots, 25$$

where $chol_i$ refers to the cholesterol level, bmi_i to the body mass index, age_i^{40-59} and age_i^{60+} are indicator variables of age for groups 40-59 and 60+, respectively, and ε_i is a Gaussian error term with zero mean and precision τ.

Note that the missing values of bmi_i are obtained from the imputation model based on linear regression discussed above using as predictors variables age_i^{40-59} and age_i^{60+}. The imputation model is specified as

$$bmi_i = \beta_{I0} + \beta_{I1} age_i^{40-59} + \beta_{I2} age_i^{60+} + \varepsilon_{Ii},\ i \in \mathcal{I}.$$

Here, $\mathcal{I}$ represents the set of indices of the observations with missing values of body mass index. Parameters β_{I0}, β_{I1}, β_{I2} represent the intercept and the covariate coefficients used in the imputation model, and ε_{Ii} is a Gaussian error with zero mean and precision τ_I. Note that all the parameters in the imputation model are mainly informed from the observed values of the body mass index and age, and their prior distributions. Because the imputation model is part of the joint model there is also feedback from all the other parts of the model when estimating the imputation model parameters and the imputed values of body mass index.

A logistic regression is used for the missingness mechanism of bmi_i under MAR or MNAR. For MAR we assumed an intercept plus the covariate of age group, while for MNAR we assumed an intercept plus the covariate of bmi_i (that includes the imputed values). For simplicity, the model with both covariates can be represented as

$$\begin{aligned} M_i &\sim Bernoulli(p_i),\ i = 1 \ldots, 25 \\ logit(p_i) &= \gamma_0 + \gamma_1 age_i^{40-59} + \gamma_2 age_i^{60+} + \delta bmi_i \end{aligned} \tag{5}$$

where M_i is a missingness indicator for bmi_i (0 for observed and 1 for missing).

Finaly, the priors for the coefficients of the fixed effects are independent Normal distributions with zero mean and precision 0.001. For the precision parameters, a Gamma with parameters 0.01 and 0.01 is used to provide a vague prior. All parameters are considered to be independent a priori.

Note that the model for analysis and the imputation model are the same for the three missingness scenarios (i.e., MCAR, MAR and MNAR). However, the missingness models differ to include different terms to accomodate the different missingness mechanisms; see Table 1 to assess which terms are included in each missingness model.

Table 1. *Posterior mean (and standard deviation) of the parameters from the joint models in the* `nhanes2` *dataset.*

Model	Parameter	Missingness mechanism in the model		
		MCAR	MAR	MNAR
Analysis	β_0	-4.084 (1.209)	-4.233 (0.816)	-4.864 (1.247)
	β_1	1.145 (0.421)	1.154 (0.398)	1.229 (0.447)
	β_2	1.866 (0.541)	1.879 (0.501)	1.940 (0.580)
	β_3	0.111 (0.049)	0.145 (0.044)	0.156 (0.044)
	τ	2.219 (0.786)	2.568 (1.312)	2.620 (1.169)
Imputation	β_{I0}	31.195 (1.569)	30.046 (1.515)	30.401 (1.296)
	β_{I1}	-5.902 (1.985)	-5.204 (2.316)	-4.711 (1.742)
	β_{I2}	-7.395 (1.733)	-5.561 (2.372)	-6.153 (2.126)
	τ_I	0.058 (0.027)	0.073 (0.023)	0.096 (0.030)
Missingness	γ_0	–	-0.337 (0.585)	-4.633 (4.892)
	γ_1	–	1.879 (0.501)	–
	γ_2	–	-0.377 (1.044)	–
	δ	–	–	0.092 (0.167)

Table 1 also shows the different estimates for all the models considered. Regarding the Gaussian analysis model, it seems that all three covariates included in the model play a significant role when explaining cholesterol level. In addition, point estimates are very similar across different missingness mechanisms. In the imputation model, we also observe that point estimates are very similar across missingness mechanisms. Age also plays an important role when imputing the missing values of body mass index. Finally, the different models for the missingness mechanism are not directly comparable.

Under MAR, age^{40-59} helps to explain why some values of body mass index are missing, while under MNAR the missing values do not appear to depend on their actual values as the estimate of δ is close to zero. We have not included age under MNAR in the missingness sub-model because this covariate is already used when imputing the missing values of body mass index, which is included in the linear predictor of the missingness model.

Cholesterol level seems to increase with age. In addition, the imputation models point to that body mass index seems to decrease with age. Although this is counterintuitive, we believe that is due to the general pattern observed in the dataset, which contains data on 25 people and only 13 of them have a complete record (i.e., all the values for all the covariates have been observed so that there are no missing values in the covariates).

As a final remark, it is worth noting that fitting these models took a few seconds. Hence, the sensitivity analysis could include other models than the ones presented here. See, for example, Mason et al. (2012) for a general discussion and alternative models for the sensitivity analysis. Larger datasets may take longer to run, but INLA will be able to fit these models faster than typical MCMC algorithms.

5.1.1. Imputation of categorical covariates with missing values

As we have mentioned in the description, this dataset includes an indicator of hypertensive status of the subjects. This categorical covariate also contains several missing values. To illustrate how missing values in continuous and categorical covariates can be handled at the same time we fit a model in which body mass index and hypertensive status are included. The imputation of body mass index will be done within the joint model as previously described, but the imputation of hypertension will be done using a multiple imputation approach; this means that an imputation model will be fit for hypertension, values of hypertensive status sampled from this model and used to fill the gaps in the original dataset. This will provide a number of complete datasets to which the analysis model will be fit; then the results will be pooled to obtain final estimates using Bayesian model averaging with equal weights Gómez-Rubio, Bivand and Rue (2020).

The analysis model becomes:

$$chol_i = \beta_0 + \beta_1 age_i^{40-59} + \beta_2 age_i^{60+} + \beta_3 bmi_i + \beta_4 hyp_i + \varepsilon_i,\ i = 1, \ldots, 25.$$

For simplicity, the missingness mechanism will not be assessed now. This implies assuming MCAR, but we have already seen that the model estimates will be close to model fit under MAR and MNAR for the case of body mass index.

The imputation model for hypertensive status (hyp_i) will be a multinomial model fit using the multinomial-Poisson transformation (Baker, 1994). This will provide estimates of the posterior probabilities of being hypertensive given the age group, which will be used to impute the missing values according to the age group of the patient. These posterior probabilities are shown in Table 2. Note that in this particular case a logistic regression would have been enough, but we have preferred to use the multinomial-Poisson transformation because it is a more general approach for the case of more than two categories.

Table 2. *Posterior probabilities of being hypertensive for the different age groups.*

	Age group		
Hypertensive	20-39	40-59	60+
Yes	1.00	0.66	0.49
No	0.00	0.34	0.51

We have drawn 100 samples to fill in the missing values of the hypertensive status, so that 100 different completed datasets have been used to fit the model. The resulting models have been pooled to obtained the posterior marginals of the model parameters using Bayesian model averaging with equal weights (Gómez-Rubio et al., 2020). These are shown in Table 3.

Table 3. *Estimates of the model parameters using multiple imputation on body mass index and hypertensive status.*

Analysis model	
Parameter	Estimate
β_0	-4.981 (1.166)
β_1	1.208 (0.518)
β_2	1.985 (0.635)
β_3	0.134 (0.072)
β_4	0.027 (0.566)
τ	1.965 (0.994)
Imputation model for bmi_i	
β_{I0}	29.612 (1.474)
β_{I1}	-3.899 (2.114)
β_{I2}	-6.116 (2.337)
τ_I	0.092 (0.034)

As expected, the estimates of the coefficients of age are close to the ones in the previous models. The coefficient of hypertensive status is close to zero, which indicates no association between cholesterol level and hypertensive status. Furthermore, the imputation model for body mass index based on a linear regression on age provides similar estimates to the imputation models fit previously and with similar effects of age on body mass index.

5.2. Simulation study: imputation of correlated data

The second example that we present is a simulation study based on the North Carolina Sudden Infant Death Syndrome (SIDS) dataset. It records several variables, which include the number of sudden infant deaths per county in the period 1974-78 (O_i), the total number of births (N_i), as well as the number of non-white births (NW_i). The expected number of cases in each county (E_i) can be obtained using internal standardization, so that the standardized mortality ratio (SMR) can be computed as O_i/E_i. Furthermore, several authors (see, for example, Cressie, 2015) have described the strong spatial pattern in the data, in the relative risk (estimated using the SMR, for example) and its correlation with the proportion of non-white births.

The model of interest to be fit is simply a Poisson regression, as follows:

$$O_i \sim Po(\mu_i); \mu_i = E_i\theta_i,\ i = 1,\ldots,100,$$

$$\log(\theta_i) = \beta_0 + \beta_1\ nwp_i.$$

Here, covariate nwp_i is the logit of the proportion of non-white births (NW_i), so that it is not bounded, that has been re-centered and re-scaled. This derived covariate has still a strong spatial pattern and a high correlation with the SMR.

Figure 1 shows the SMR for the period 1974-78 and the transformed proportion of non-white births (nwp_i). The SMR shows some areas of high risk and a strong correlation with the proportion of non-white births. Hence, this covariate can be useful when building models to explain the spatial variation of SIDS in North Carolina.

The simulation study will remove 5%, 10%, 15%, 30% and 50% of the covariate values (i.e., proportion of non-white births) using MCAR and MNAR mechanisms. Note that MAR can be regarded as an extension to MCAR that considers other observed covariates in the linear predictor of the logistic regression in the imputation model. Although MAR may seem more reasonable, it is simply a matter of including other covariates in the linear predictor of the missingness model so it is computationally feasible but it adds little to the comparison. This is why we have not considered it.

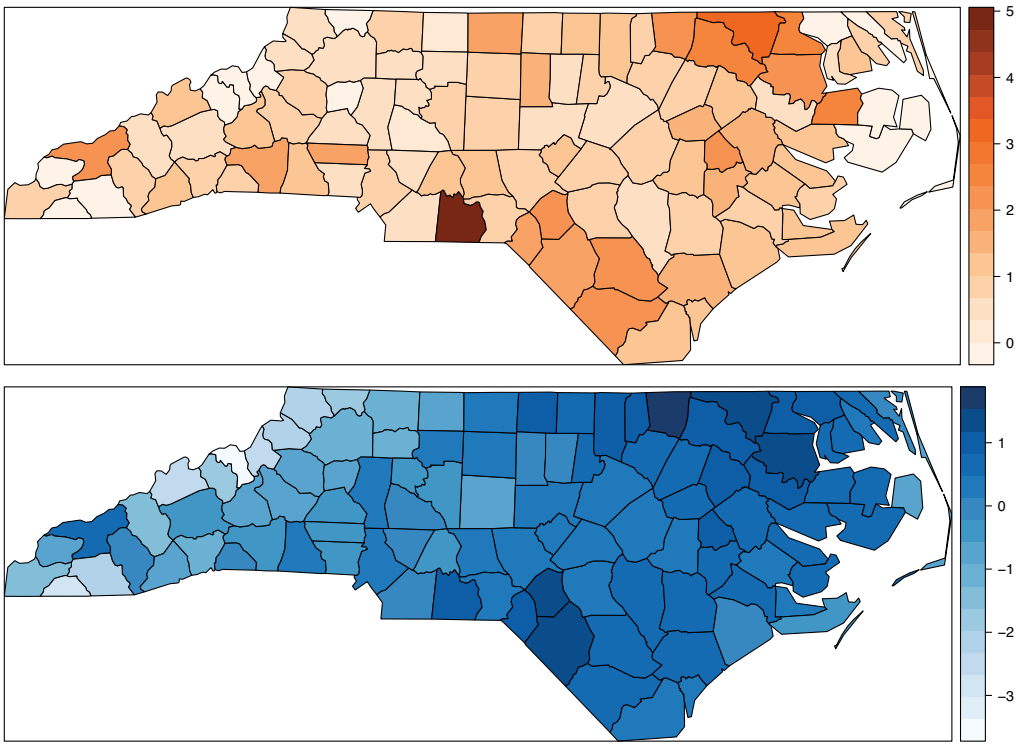

Figure 1. *Standardized mortality ratio (SMR, top) and proportion of non-white births (bottom) in North Carolina in the period 1974-78.*

The missing observations will be nested across the five scenarios, i.e., the observations removed in the 10% scenario will also be removed in the 15% scenario and so on. Furthermore, the probability of being missing under the MNAR mechanism p_i is

$$logit(p_i) = \alpha_M + 5x_i$$

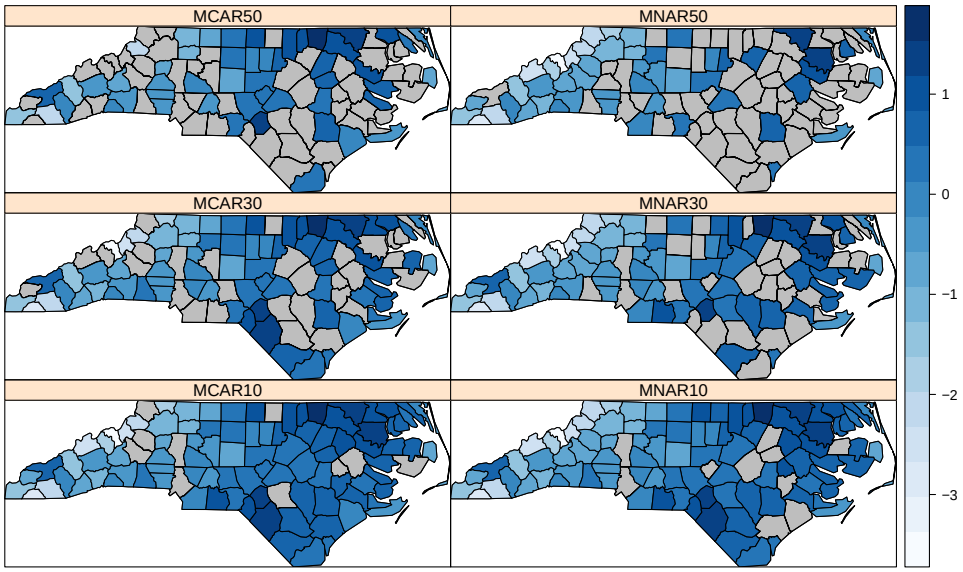

Figure 2. *Missing observations (in grey) of the proportion of non-white births.*

where α_M is set as the logit of 0.5 and x_i represents the value of the covariate with missing values.

This simulation is intended to compare mild to severe missingness under five different scenarios for MCAR and MNAR. Models will be fit assuming MCAR and MNAR missingness, so that we fit 20 models in total. Under MCAR, we only fit the analysis and imputation model. Under MNAR, in addition we will assess whether the joint approach including the missingness mechanism is able to capture the type of missingness.

Figure 2 shows the missing values of the proportion of non-white births for three of the scenarios considered in this simulation study. As it can be seen, when the percentage of missing values is 50% under MNAR missing values concentrate in the counties with high values of the covariate.

In addition, the imputation model proposed is based on the conditional autoregressive specification presented in Section 3.3, so that imputation is included within the main model. This imputation model will have the following parameters: τ is the precision of the CAR specification, ρ the spatial autocorrelation and α the mean value of the covariates.

Finally, a logistic regression on the missingness variable M_i (0 for observed and 1 for missing) is used to model the missingness mechanism (under MNAR):

$$\begin{aligned} M_i &\sim Bernoulli(p_i);\ i = 1,\ldots,100 \\ logit(p_i) &= \gamma_0 + \gamma_1 nwp_i \end{aligned} \tag{6}$$

Note that the imputed values appear both in the Poisson regression and the sub-model on the missingness mechanism. Non-zero values of γ_1 indicate that the probability of being missing depends on the actual values.

Table 4 summarises the models fit to the data under MCAR. Here, an imputation sub-model for the covariate has been included but not a joint model for the missingness as under MCAR it is not necessary. In general, there are not large differences between the different models fit to the datasets regarding percentage of missing values and type of actual missingness. However, these differences become larger as the proportion of missing values increases, which was to be expected. These differences are noticeable for the case of 50% of missing values both under MCAR and MNAR.

The estimates of the imputation models are quite similar as well, across missingness type in the data and proportion of missing values. However, some differences are observed for 30% and 50% of missing values. In particular, the estimates of α differ.

Table 5 summarises the (joint) models fit to the data considering a MNAR scenario. This includes the model fit to the complete dataset, and the binomial sub-model in the joint model to assess the missingness mechanism. First of all, the posterior distribution of γ_1 helps to determine the missingness mechanism. Its posterior estimate is very close to zero under MCAR, while it is above zero under MNAR (but for the case of 5% of missing values). It is worth stating that it is possible to assess this now because these are simulated data and the true missingness mechanism is known.

Regarding the imputation model, the estimates are very similar across scenarios. Finally, the estimates of the parameters in the Poisson model are in general very close to the model fit to the full dataset.

It is worth noting that under MNAR with 50% of missing observations the point estimates of the parameters in the Poisson sub-model show the largest departure from the model fit to the full dataset. This is probably due to the fact that the imputation model is not able to fully recover the values of the covariates as missing values tend to have high values and there is not enough information in the observed values as to recover this pattern.

To sum up, imputation models behave as expected and provide a good performance in all cases. Most importantly, the joint model is able to identify between MCAR and MNAR situations as well as imputing the covariates and fit the model of interest to the data. Again, this is possible now because the missingness mechanism is known but in real applications we would propose different models and conduct a sensitivity analysis.

When the models fit under MCAR (Table 4) and under MNAR (Table 5) are compared, it should be mentioned that when data under MCAR are analysed both models produce very similar results because the missingness mechanism is, in fact, independent of the observed data. For the analysis of the data simulated under MNAR, differences can be observed because now the missingness mechanism depends on the covariate (including the imputed data) and the estimates of the parameters in the imputation sub-model are different.

Table 4. *Posterior mean (and standard deviation) of the model parameters under MCAR.*

		Model under MCAR						
		Poisson		Imputation			Missingness	
Missingness	% missing	β_0	β_1	τ	ρ	α	γ_0	γ_1
–	0	-0.141 (0.046)	0.524 (0.068)	–	–	–	–	–
MCAR	5	-0.126 (0.047)	0.518 (0.068)	2.129 (0.305)	0.977 (0.022)	-0.211 (0.162)	–	–
MCAR	10	-0.114 (0.047)	0.496 (0.069)	2.076 (0.301)	0.976 (0.024)	-0.215 (0.165)	–	–
MCAR	15	-0.120 (0.048)	0.504 (0.067)	1.915 (0.294)	0.973 (0.027)	-0.234 (0.175)	–	–
MCAR	30	-0.099 (0.049)	0.507 (0.065)	1.776 (0.295)	0.960 (0.039)	-0.175 (0.183)	–	–
MCAR	50	-0.077 (0.051)	0.518 (0.070)	2.461 (0.481)	0.957 (0.044)	0.034 (0.169)	–	–
MNAR	5	-0.131 (0.045)	0.506 (0.067)	2.040 (0.292)	0.977 (0.022)	-0.236 (0.166)	–	–
MNAR	10	-0.138 (0.048)	0.506 (0.068)	1.991 (0.288)	0.976 (0.023)	-0.220 (0.167)	–	–
MNAR	15	-0.110 (0.048)	0.495 (0.068)	1.966 (0.289)	0.976 (0.024)	-0.238 (0.170)	–	–
MNAR	30	-0.105 (0.050)	0.453 (0.070)	1.827 (0.291)	0.975 (0.025)	-0.342 (0.189)	–	–
MNAR	50	-0.064 (0.055)	0.419 (0.061)	1.421 (0.279)	0.964 (0.037)	-0.423 (0.226)	–	–

Table 5. *Posterior mean (and standard deviation) of the model parameters under MNAR.*

		Model under MNAR						
		Poisson		Imputation			Missingness	
Missingness	% missing	β_0	β_1	τ	ρ	α	γ_0	γ_1
–	0	-0.141 (0.046)	0.524 (0.068)	–	–	–	–	–
MCAR	5	-0.121 (0.047)	0.512 (0.068)	2.120 (0.305)	0.977 (0.022)	-0.217 (0.163)	-3.218 (0.565)	-0.514 (0.465)
MCAR	10	-0.111 (0.048)	0.494 (0.069)	2.073 (0.301)	0.975 (0.024)	-0.216 (0.165)	-2.271 (0.349)	-0.074 (0.392)
MCAR	15	-0.127 (0.049)	0.505 (0.067)	1.903 (0.293)	0.972 (0.027)	-0.218 (0.176)	-1.821 (0.309)	0.359 (0.396)
MCAR	30	-0.110 (0.050)	0.507 (0.065)	1.768 (0.294)	0.960 (0.039)	-0.141 (0.187)	-0.896 (0.232)	0.339 (0.309)
MCAR	50	-0.079 (0.054)	0.518 (0.070)	2.458 (0.480)	0.956 (0.044)	0.040 (0.176)	-0.014 (0.203)	0.038 (0.307)
MNAR	5	-0.132 (0.045)	0.502 (0.068)	2.046 (0.293)	0.976 (0.023)	-0.236 (0.165)	-3.286 (0.720)	0.810 (0.795)
MNAR	10	-0.153 (0.049)	0.486 (0.071)	1.964 (0.287)	0.977 (0.022)	-0.225 (0.170)	-2.947 (0.849)	1.661 (0.828)
MNAR	15	-0.133 (0.049)	0.481 (0.069)	1.928 (0.287)	0.977 (0.023)	-0.227 (0.173)	-2.225 (0.529)	1.306 (0.592)
MNAR	30	-0.152 (0.052)	0.423 (0.069)	1.688 (0.285)	0.976 (0.024)	-0.190 (0.200)	-1.385 (0.450)	1.477 (0.492)
MNAR	50	-0.172 (0.060)	0.380 (0.060)	1.230 (0.266)	0.969 (0.032)	-0.093 (0.253)	-0.303 (0.351)	1.576 (0.434)

Finally, we have included the posterior distributions of some imputed values of the covariate in Figure 3. In particular, we have considered the dataset with 50% missing values under MNAR and taken nine counties with missing values that have missing values also in the simulated data under MCAR. This produces a set of counties with a wide variety in the posterior marginals of the imputed values. The posterior marginals shown are for the imputation model under MCAR in Table 4 (dashed line) and the imputation model under MNAR in Table 5 (dotted line). The vertical solid line shows the actual value of the missing covariate. Furthermore, we have kept the same axes scale in all plots so that differences are appreciated better.

In general, both marginals are close in all cases. Under MNAR (dotted lines), the posterior mode seems to be closer to the actual value for most of the counties in the plot. This should not be surprising as this is the actual missingness mechanism in the data.

As the counties considered here are also present in the case in which the missingness mechanims is MCAR, it could be possible to check what happens between models

that assumed MCAR and MNAR when the actual missingness is MCAR. In this case, the posterior marginals of the missing values (assuming MCAR and MNAR) look the same for each county because accounting for the missingness model does not affect the model estimates. This shows that handling imputation of missing values with INLA is an interesting way to conduct sensitivity analysis.

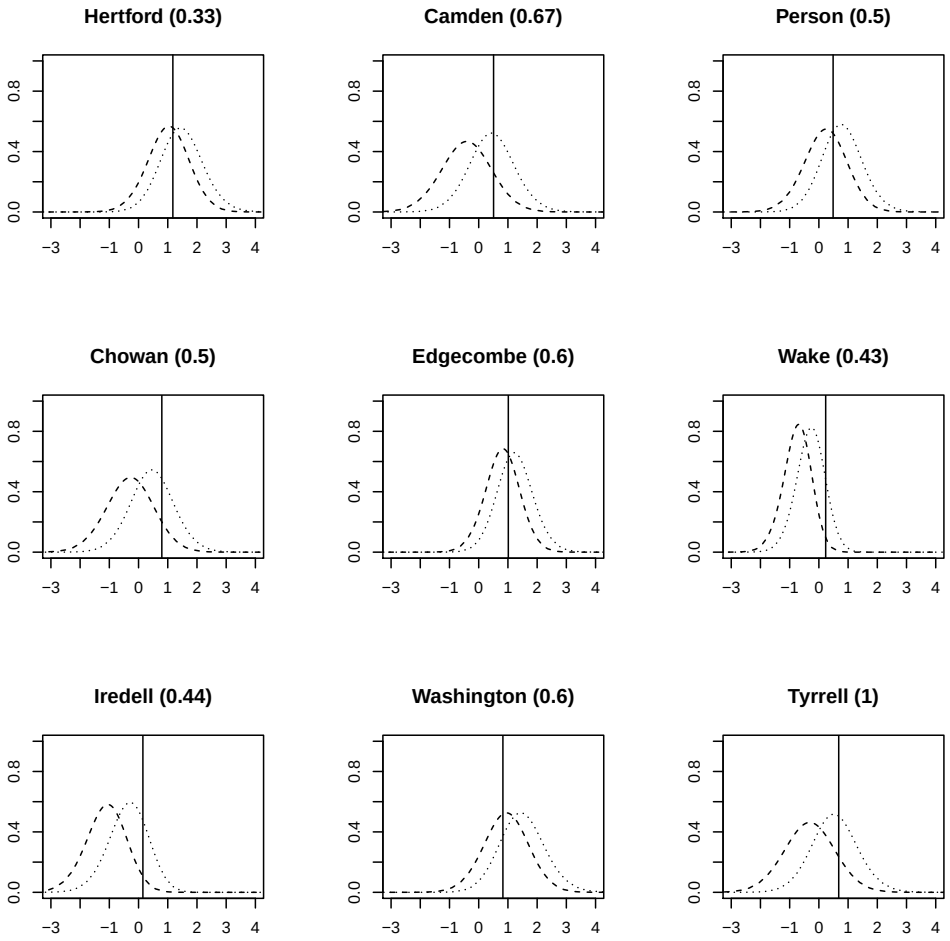

Figure 3. *Posterior marginal distributions of some of the imputed values for missingness of 50% under MNAR. The lines represent the actual value (solid vertical line), the posterior marginal from the MCAR model (dashed line) and the posterior marginal from the MNAR model (dotted line). The value between parenthesis corresponds to the proportion of missing values in the neighbour counties.*

6. Discussion

This paper shows how the general problem of dealing with missing observations in the covariates and performing multiple imputation under different missingness mechanisms can be recast within the framework of latent Gaussian Markov random field models. This has the main advantage that models expressed as latent GMRFs can be fit through INLA, making inference fast. Furthermore, this fills an important gap in the INLA methodology as now models with missing values in the covariates can be easily fit.

Imputation models for the covariates can also take many different forms when defined as GMRFs. In this work we have only considered a linear regression model and spatially correlated model for imputation, but other similar imputation models could be easily developed. For example, these could tackle missing observations in longitudinal data or time series. Furthermore, the methods proposed can be extended to consider imputation of more than one covariate at the same time by relying on multivariate Gaussian models.

The implementation of the multiple imputation models take the form of new latent effects for the `R-INLA` package and they are available within the `MIINLA` package for the `R` programming language. These new latent effects have been developed using the `rgeneric` framework for latent effects development within the `R-INLA` package. Nonetheless, this approach could be implemented in any other software packages for Bayesian inference.

Although we have focused on imputation of continuous covariates, missing values in categorical covariates can also be handled. However, as stated in the paper, this case does not fit within the paradigm of latent GMRF models easily. However, INLA can be used to propose an imputation model for the missing categorical data and to fit the model of interest to these imputed datasets. The fitted models can then be combined to account for the uncertainty of the imputed values in the estimation of the model parameters using Bayesian model averaging.

When the missing values of the categorical covariates index a latent effect the imputation of missing values becomes more complex. This is the case, for example, when random effects are estimated for different groups in the data using multilevel models. However, this scenario could also be handled using the multiple imputation methods described in this paper.

In addition to handling and imputing missing values, this new framework allows us to consider the missingness mechanism using a joint model fit within the INLA methodology. Hence, the analysis of data with missing observations can now be completely carried out within the INLA framework.

Sensitivity analysis on the missingness mechanism, required when it is not ignorable, can benefit from the the computational speed of the INLA method. First of all, models are fit faster than with typical MCMC methods, which helps to define the scenarios to test. Secondly, more scenarios can be tested as the time required to fit the models is reduced.

Acknowledgements

V. Gómez-Rubio has been supported by grants MTM2016-77501-P and PID2019-106 341GB-I00 from the Spanish Ministry of Economy and Competitiveness co-financed with FEDER funds, grant SBPLY/17/180501/000491 and SBPLY/21/180501/000241 funded by Consejería de Educación, Cultura y Deportes (JCCM, Spain) and FEDER. Marta Blangiardo acknowledges partial support through the grant R01HD092580 funded by the National Institute of Health and from the MRC Centre for Environment and Health, which is currently funded by the Medical Research Council (MR/S019669/1).

References

Baker, S. G. (1994). The multinomial-Poisson transformation. *The Statistician*, 43(4):495–504.

Barber, X., Conesa, D., Lladosa, S., and Lòpez-Quílez, A. (2016). Modelling the presence of disease under spatial misalignment using Bayesian latent Gaussian models. *Geospatial Health*, 11(1):1–10.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. John Wiley & Sons, Ltd, Chichester, UK.

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

Carpenter, J. R. and Kenward, M. G. (2012). *Multiple Imputation and its Application*. John Wiley & Sons, Ltd, Chichester, UK.

Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society: Series A*, 169:571–584.

Carpenter, J. R., Kenward, M. G., and White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random - a weighting approach. *Statistical Methods in Medical Research*, 16:259–275.

Cressie, N. (2015). *Statistics for Spatial Data*. John Wiley & Sons, Inc., Hoboken, NJ, revised edition.

Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press, New York, NY.

Erler, N. S., Rizopoulos, D., Rosmalen, J. v., Jaddoe, V. W. V., Franco, O. H., and Lesaffre, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, 35(17):2955–2974.

Forlani, C., Bhatt, S., Cameletti, M., Krainski, E., and Blangiardo, M. (2020). A joint Bayesian space-time model to integrate spatially misaligned air pollution data in R-INLA. *Environmetrics*, 31(8):e2644.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/ Hierarchical Models*. Cambridge University Press, New York.

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. Chapman & Hall/CRC, Boca Raton, FL.

Gómez-Rubio, V., Bivand, R. S., and Rue, H. (2020). Bayesian model averaging with the integrated nested Laplace approximation. *Econometrics*, 8(2):23.

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

Held, L. and Rue, H. (2010). Conditional and intrinsic autoregressions. In Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P., editors, *Handbook of Spatial Statistics*, chapter 13, pages 201–216. Chapman & Hall.

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*, 73(4):423–498.

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

Little, R. J. A. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons, Inc., Hoboken, NJ, 3rd edition.

Martino, S. and Riebler, A. (2019). Integrated nested Laplace approximations (INLA). arXiv:1907.01248 [stat.CO].

Mason, A., Richardson, S., Plewis, I., and Best, N. (2012). Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. *Journal of Official Statistics*, 28(2):279–302.

Mason, A. J. (2009). *Bayesian methods for modelling non-random missing data mechanisms in longitudinal studies*. Ph.D., Imperial College London.

Nakagawa, S. (2015). Missing data: mechanisms, methods, and messages. In Fox, G. A., Negrete-Yankelevich, S., and Sosa, V. J., editors, *Ecological Statistics: Comtemporary Theory and Practice*, chapter 4, pages 81–105. Oxford University Press, Oxford.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383.

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

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91(434):473–489.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton, FL.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 2(71):1–35.

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(1):395–421.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295.

Trivellore, R. (2015). *Missing Data Analysis in Practice*. Chapman & Hall/CRC, Boca Raton, FL.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, FL.

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

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.