pagepress

Geospatial Health
Geospatial Health 2026; 21:1511
doi:10.4081/gh.2026.1511

Bayesian hierarchical spatial models for disease mapping in the presence of missing covariates

Sami Ullah, Tianfa Xie

School of Mathematics, Statistics and Mechanics, Beijing University of Technology, Beijing, China

Abstract

Bayesian spatial models for disease mapping, such as Besag-York-Mollié (BYM) models, are widely used to model disease counts while accounting for spatial dependence. However, these models are not equipped to handle missing covariate values. Covariates are often partially observed, yet these models require separate imputation that ignores imputation uncertainty. We extend this Bayesian framework for disease mapping to accommodate missing covariates while modelling the disease counts. Missing covariate values are treated as unknown parameters that are estimated simultaneously with the other model's parameters within the same Bayesian model. The proposed model was evaluated across increasing missing proportions using two real datasets: the Scottish lip cancer dataset with a single covariate and the North Carolina SIDS dataset with two covariates. The results show that the proposed model performs well with a missing proportion of up to 30% in a single covariate or a cumulative 30% across multiple covariates.

Key words: Bayesian spatial models, missing covariates, BYM models, disease mapping.

Correspondence: Tianfa Xie, School of Mathematics, Statistics and Mechanics, Beijing University of Technology, Beijing 100124, China.
Tel.: +8613693580632; E-mail: xietf@bjut.edu.cn

Introduction

Spatial statistical approaches and spatial smoothing models have gained significant attention in public health and ecology. It is recommended to borrow strength from nearby locations to obtain more accurate region-specific estimates rather than directly inspecting crude rates due to inherent sampling variability. The state-of-the-art model in this domain is the Besag, York, and Mollié (BYM) model (Besag *et al.*, 1991). The BYM model is a log-linear Poisson model developed for disease risk mapping, which includes a spatially structured component to improve log risk estimation by borrowing strength from neighbouring areas and an unstructured component to account for regional heterogeneity. The BYM approach uses the Poisson model to estimate the log relative risk η_i for region i (i = 1, 2, …N) based on the observed disease counts. It specifies:

$$\eta_i = \alpha + X_i^T \beta + \theta_i \sigma_\theta + \phi_i \sigma_\phi \quad \text{Eq. 1}$$

where α is the intercept of the logistic model; $_{Xi}$ a vector of covariates for the i-th spatial region; β a vector of coefficients; ϕ_i the spatial random effect, which captures spatial dependencies across neighbouring regions; and θ_i the heterogeneous random effects, accounting for the unstructured variation across the regions, with σ_ϕ and σ_θ the Standard Deviations (SDs) of the spatial and heterogeneous random effects, respectively. This model has been widely used in disease mapping and ecology (Slater *et al.*, 2022; Cifo *et al.*, 2024; Leveau *et al.*, 2024; Bhakkan-Mambir *et al.*, 2025; Haque *et al.*, 2025).

Several modifications and alternative approaches have been proposed; see, for example, the Leroux model (2000) and the Dean model (2001). The suitability and behaviour of various latent spatial models have been compared using penalized quasi-likelihood, empirical Bayes or a full Bayesian context (Best *et al.*, 2005; Wakefield, 2007; Lee, 2011; MacNab, 2011). Simpson *et al.* (2017) introduced a new BYM parameterization known as the BYM2 model, which simplifies prior specifications by aligning them with the model structure, ensuring that all parameters are interpretable. This new parameterization follows the framework of the Leroux model, placing a single precision (σ) on the combined components (spatial and non-spatial) and a mixing parameter (ρ) for spatial and non-spatial variation. To ensure that σ is the SD of the combined components, the variances of the spatial effect ϕi_i and the non-spatial effects θ_i are scaled to be approximately 1, with a scaling factor computed from the neighbourhood graph of the study area. Furthermore, Morris *et al.* (2019) offered an implementation of the BYM2 model in Stan (a probabilistic programming platform for full Bayesian inference using Hamiltonian Monte Carlo, HMC). However, these models are not equipped to handle missing covariate values. Covariate values are often unknown for some spatial units, and the current BYM models require separate imputations, which these models consider as fixed values, thus ignoring the uncertainty associated with these imputations. This leads to overconfident estimation, distorted spatial structure, and spurious findings in Bayesian spatial models. Therefore, a joint Bayesian model for the missing covariates and the outcome variable should be specified to fully account for the uncertainty in imputed values.

The challenges of missing covariates have been addressed in a spatial Gaussian regression model for a continuous outcome (Ma *et al.*, 2021). This model effectively handles missing covariates through an autoregressive structure. However, it is not directly

applicable to disease count data, which typically require Poisson models and different spatial specifications as in BYM models. This study aims to extend the Poisson-based BYM modelling framework for disease mapping to accommodate spatially dependent missing covariates while modelling disease counts. To borrow strength from neighbouring spatial units, missing covariate values were modelled through a spatial structure that considers a covariate mean and a scaled spatial random effect. We further evaluated the proposed joint model using two real-world datasets: the Scottish lip cancer dataset with a single covariate and the North Carolina sudden infant death syndrome (SIDS) dataset with two covariates, by repeatedly masking different proportions of covariate values using a random mechanism. Model performance was assessed in terms of its ability to recover the parameters of interest relative to the complete-data BYM2 model, as well as prediction accuracy and uncertainty quantification.

Model specification

Suppose a study area is subdivided into N spatial regions indexed by $i = 1, 2, \ldots, N$ with a known spatial adjacency structure. For each spatial region, let Y_i be the observed disease count, E_i the expected count and $X_1,\ldots, X_{pare}$ continuous covariates that are partially observed, such that each covariate has a known value for the region i∈O and unknown for the region i∈M, where O ∪ M={1,…,N}. We develop a joint Bayesian spatial model that simultaneously models unknown covariate values and the observed disease count Y_i. This joint model consists of two components: modelling the observed disease counts Y_i and imputation of unknown covariate values.

Modelling disease counts

The disease counts Y_i are modelled with a log-linear Poisson model as follows:

$$Y_i|\lambda_i \sim \text{Poisson}(\mu_i),$$
$$\log(\mu_i) = \log(E_i) + \beta_0 + \sum_{j=1}^{p}\beta_j X_{ji} + u_i \quad \text{Eq. 2}$$

where β_0 denotes the intercept; βj j-th covariate's effect; and u_i an area-level latent random effect.

Following the BYM2 model's parameterization, the latent random effect u_i is decomposed into spatially structured and unstructured components as follows:

$$u_i = \sigma_u\left((\sqrt{\rho/s})\phi_i^Y + (\sqrt{1-\rho})\theta_i\right) \quad \text{Eq. 3}$$

where ρ∈(0,1) is the proportion of variation that comes from the spatially correlated error; ϕ_i^Y the spatial effect that captures spatial dependence; θ_i unstructured heterogeneity modelled as θ_i~N(0,1); σ_u the overall SD for the combined error terms; and s is the scaling factor computed from the neighbourhood graph as in the BYM model, ensuring var(ϕ_i)≅1.

Following the BYM2 parameterization, the spatial effect ϕ_i^Y were modelled with the Intrinsic Conditional Autoregressive (ICAR) model defined by the adjacency structure as follows:

$$p(\phi^Y) \propto exp\left(-\tfrac{1}{2}\textstyle\sum_{i\sim k}(\phi_i^Y - \phi_k^Y)^2\right) \quad \text{Eq. 4}$$

with a soft sum-to-zero constraint for identifiability:

$$\textstyle\sum_{i=1}^{N}\phi_i^Y \sim N(0, 0.001N).$$

Modelling unknown covariate values

To model unknown covariate values, we used the BYM2 spatial structure, which captures covariate-specific spatial effects through an ICAR model, separating the spatially structured effect from the unstructured ones. For the partially observed covariate X_j(j=1,…, p), we define X_j={$X_j1,\ldots,XjN$} that includes both observed and imputed values as follows:

$$X_{ji} = \begin{cases} X_{ji}^{obs} \; for\; i \in O \\ X_{ji}^{miss} \; for\; i \in M \end{cases} \quad \text{Eq. 5}$$

where X_j^{obs} denotes the known values: X_j^{miss} denotes the unknown/missing values of j-th covariate. Missing values are treated as unknown parameters and inferred jointly with other model components.

The covariate values (both observed and missing) are assumed to arise from a common latent spatial process.

$$X_{ji}^{obs} \sim N\left(X_{ji}^{spat}, \sigma_j^2\right),\; i \in O$$
$$X_{ji}^{miss} \sim N\left(X_{ji}^{spat}, \sigma_j^2\right),\; i \in M$$

where the latent spatial covariate surface is defined as:

$$X_{ji}^{spat} = \mu_{Xj} + \sigma_{Xj}\left((\sqrt{\rho_j/s})\phi_i^{Xj} + (\sqrt{1-\rho_j})\theta_i^j\right) \quad \text{Eq. 6}$$

where μ_{Xj} is the overall mean of the j-th partially observed covariate; ϕ_i^{Xj} the covariate-specific spatial effect of the j-th covariate for region I; σ_{Xj} regulates the magnitude of spatial variation in the j-th covariate values; ρ_j the proportion of variation in the j-th partially observed covariate that comes from the spatially correlated error; s the scaling factor; and θ_i^j the unstructured heterogeneity in j-th partially observed covariate modelled as θ_i^j~N(0,1).

The j-th covariate-specific spatial effects ϕ_i^{Xj} follow an ICAR model as:

$$p(\phi^{Xj}) \propto exp\left\{-\tfrac{1}{2}\textstyle\sum_{i\sim k}(\phi_i^{Xj} - \phi_k^{Xj})^2\right\} \quad \text{Eq. 7}$$

with a soft sum-to- zero constraint for identifiability:

$$\textstyle\sum_{i=1}^{N}\phi_i^{Xj} \sim N(0, 0.001N).$$

It should be noted that the covariate spatial effects for each partially observed covariate are modelled as distinct latent processes to avoid spatial confounding.

Though we only impute the unknown covariate values X_j^{miss}, the known covariate values X_j^{obs} are modelled to identify the latent spatial covariate process to confirm that both known and unknown values arise from the same data-generating process. This approach allows the model to concurrently learn from the known covariate values while utilizing the spatial structure to inform imputation.

Prior distributions

Weakly informative priors are assigned to the model's parameters to ensure identifiability and allow the data to dominate inference. The parameters (β_o, β_j) follow a normal prior with mean zero and a variance chosen to reflect reasonable effect sizes while providing regularization. The μ_{Xj} follows a data-informed normal prior centered at the mean of the observed covariate values with a variance sufficiently large to avoid overconfidence. The scale parameters for the covariate and the outcome variable (σ_X, σ_u) follow a half-normal prior with the scale chosen based on the expected variability in each parameter. The measurement error σ_j for modelling both the observed and the missing covariate values follows a half-normal prior with a small scale, allowing for minimal deviation while being adaptable to the data. For mixing parameters (ρ,ρ_j), we use a symmetric beta distribution with both shape parameters equal to 0.5 (equivalent to the Jeffreys prior as used in the BYM model) to learn the degree of spatial structure from the data.

Model implementation and computation

The model is implemented in 'Rstan' within R version 4.4.2 using HMC and the No-U-Turn Sampler (NUTS). Convergence is assessed using the potential scale reduction factor *(Rhat*) and effective sample size (*n_eff*). *Rha*t (Gelman-Rubin convergence diagnostic) is a statistic used in Markov Chain Monte Carlo (MCMC) analysis for comparing between-chain variance to within-chain variance; a value ≤1.01 indicates convergence (Vehtari *et al.*, 2021). This approach allows us to approximate the posterior distribution for each parameter and subsequently derive estimates such as the posterior mean, median, and 95% Credible Intervals (CIs). These results provide both point estimates and uncertainty quantification, which are essential for interpreting parameter dependencies and assessing the model fit.

Performance evaluation

To evaluate the performance of our proposed method under realistic scenarios, we designed stress tests using two real datasets: a Scottish lip cancer dataset with a single covariate and the North Carolina SIDS dataset with two covariates.

The Scottish lip cancer dataset

In this section, we evaluate the proposed model using the Scottish lip cancer dataset (Clayton and Kaldor, 1987). This dataset contains the observed and expected number of lip cancer cases in males across 56 spatial units for the period 1975–1980, along with the covariate: percentage of the workforce employed in agriculture, fishing or forestry. This is considered a foundational dataset in spatial epidemiology and is often used to test spatial models. The BYM2 model has analyzed this dataset in (Morris, 2019), and we use those results as a baseline. Following the standard practice in (Morris, 2019), we prepared the adjacency structure for Stan using the 'mungeCARdata4stan' function from the Stan case study (Morris, 2019) and scaled the covariate by a factor of 0.1 to improve numerical stability.

To evaluate the model under different missing proportions, we repeatedly masked 10%, 20%, 30% and 40% of covariate values using a random mechanism. In this way, we got four incomplete datasets, each with a set of missing regions such that at 10% missing proportion, = {13, 14, 33, 34, 55}; at 20%, = {3, 4, 8, 11, 21, 25, 29, 31, 40, 43, 45}, at 30%, = {3, 4, 6, 9, 11, 13, 18, 19, 29, 35, 36, 40, 41, 51, 55, 56} and at 40%, = {2, 3, 6, 8, 11, 16, 17, 23, 28, 30, 31, 32, 33, 37, 39, 41, 42, 43, 45, 46, 47, 56}. The proposed model was then applied to these incomplete datasets.

We assigned the weakly informative priors for the model parameters as: β_0,β_1,σ_{Xj},σ_u~N(0,1); μ_{Xj}~N(X_j^{obs},5); σ_j~N(0,0.1); ρ,ρ_j~Beta(0.5,0.5). The model was implemented in RStan using 4 chains, each run for 2,000 iterations, with the first 2,000 for 'warm-up', yielding a total of 4,000 posterior draws. The R code for the proposed method is provided in Appendix A in the supplementary file. The results of the proposed model at different levels of missing covariate values are shown in Tables 1-4, presenting the posterior summaries of model parameters of interest, as well as a fitted value μ[5] for the 5th spatial unit, computed in the generated quantity block as:

$$\mu[i] = e^{(log(E_i)+\beta_0+\sum_{j=1}^{p}\beta_j X_{ji}+u_i)} \quad \text{Eq. 8}$$

Our primary goal was to evaluate the model's effectiveness in recovering the parameters of interest compared to the complete-data BYM2 model's results (Morris, 2019), as given in Table 5.

To assess the recovery of the parameters of interest, we compare the proposed model's results at different levels of missing data (Tables 1-4) against the complete-data BYM2 model's result (Table 5). The complete-data BYM2 model, implemented in 'CmdStan', serves as the baseline truth that our model should approximate. Figure 1 compares absolute deviations of the parameter estimates from the baseline at different missing proportions. While most parameter estimates exhibited a gradual increase, (the fitted value for the 5th spatial unit) deviated sharply at 30% and 40% missing proportion, followed by the fixed-effects β_0 and β_1. Conversely, spatial parameters remained relatively stable across varying missing proportions, indicating that our model borrows strength from neighbours even with substantial missing values in a covariate. These results suggest that the proposed model is efficient in recovering the parameter estimates of the complete-data BYM2 model under low-to-moderate missing proportions in a covariate. For comparing the credible intervals, it should be noted that our model was implemented in RStan, which reported 95% credible intervals. In contrast, the published results of the complete-data BYM2 model (fitted in CmdStan) reported 90% credible intervals. Therefore, we assess consistency through interval overlap. The credible intervals for all estimates in our model show substantial overlap with the 90% credible intervals of the complete-data BYM2 model at all missing levels, demonstrating that the additional variability introduced by missing data is appropriately captured. Although *Rhat* values for all parameters of interest suggested convergence, diagnostic warnings (divergent transitions, low Effective Sample Size (ESS) were observed. These diagnostic issues stem from the limited information in the smaller Scottish dataset (56 spatial units), not from model misspecification, as evidenced by the fact that these issues were not observed with the larger North Carolina dataset (100 spatial units) presented below.

The North Carolina SIDS dataset

To evaluate the performance of the proposed model in the scenarios of missing values in multiple covariates, we use the North Carolina SIDS dataset (Cressie and Chan, 1989), publicly available from the 'sf-package' in R software. This dataset contains sudden infant death Syndrome Counts (SID), total births, and area for 100 spatial units in North Carolina over the time periods 1974-1978 and 1979-1984. We used the number of SIDs deaths from 1979-1984 as an outcome variable (Y=SID79). The expected number of SIDS counts was calculated as the total births in the *i*-th spatial unit multiplied by the overall SIDS rate during 1979-1984. We considered the two covariates, the previous SID rate (X_{1i}) and pop-

ulation density (X_{1i}), calculated as:

$$X_{1i} = \frac{\text{Total SIDS cases in ith county during } 1974 - 1978}{\text{Total births in that county during } 1974 - 1978} \quad \text{Eq. 9}$$

$$X_{2i} = \frac{\text{Total births in the ith county during } 1979 - 1984}{\text{Area of that county during } 1979 - 1984} \quad \text{Eq. 10}$$

Both covariates were standardized to have a mean of zero and unit variance to ensure comparable coefficients and improved MCMC convergence. To evaluate the model under different missing proportions, we randomly masked 10%, 15%, and 20% of the

Table 1. Results at 10% missing proportion in Scottish lip cancer data.

Parameter	Mean	SE	SD	X2.5	X97.5	n_eff	Rhat
β_0	-0.22	0.00	0.13	-0.47	0.03	2057	1.00
β_1	0.38	0.00	0.14	0.10	0.65	1858	1.00
σ_u	0.52	0.00	0.08	0.37	0.71	1155	1.00
ρ	0.88	0.01	0.14	0.48	1.00	616	1.01
$\mu[5]$	13.79	0.04	3.13	8.39	20.61	5699	1.00
$\phi^Y[5]$	1.38	0.01	0.40	0.65	2.24	2067	1.00
$\theta[5]$	0.17	0.01	0.95	-1.68	2.03	6018	1.00

Mean, posterior mean; SE, Monte Carlo standard error (variability of the posterior mean estimate); SD, posterior standard deviation (measure of posterior uncertainty); X2.5, 2.5% posterior quantile (lower bound of the 95% credible interval); X97.5, 97.5% posterior quantile (upper bound of the 95% credible interval); n_eff, effective sample size (number of independent draws from the posterior); Rhat, Gelman-Rubin convergence diagnostic (a value ≤ 1.01 indicates convergence).

Table 2. Results at 20% missing proportion in Scottish lip cancer data.

Parameter	Mean	SE	SD	X2.5	X97.5	n_eff	Rhat
β_0	-0.22	0.00	0.15	-0.51	0.08	1316	1.00
β_1	0.34	0.00	0.15	0.03	0.64	1207	1.00
σ_u	0.54	0.00	0.09	0.38	0.73	1028	1.01
ρ	0.87	0.01	0.15	0.48	1.00	558	1.00
$\mu[5]$	13.58	0.07	3.15	8.07	20.53	1832	1.00
$\phi^Y[5]$	1.36	0.01	0.38	0.65	2.16	1340	1.00
$\theta[5]$	0.20	0.02	0.92	-1.58	1.93	5189	1.00

Table 3. Results at 30% missing proportion in Scottish lip cancer data.

Parameter	Mean	SE	SD	X2.5	X97.5	n_eff	Rhat
β_0	-0.28	0.00	0.15	-0.58	0.01	1910	1.00
β_1	0.43	0.00	0.15	0.12	0.72	1699	1.00
σ_u	0.52	0.00	0.09	0.36	0.71	1115	1.00
ρ	0.90	0.00	0.13	0.53	1.00	928	1.00
$\mu[5]$	13.44	0.04	3.08	8.16	20.19	5298	1.00
$\phi^Y[5]$	1.36	0.01	0.39	0.64	2.18	2136	1.00
$\theta[5]$	0.20	0.01	0.97	-1.70	2.03	7286	1.00

Table 4. Results at 40% missing proportion in Scottish lip cancer data.

Parameter	Mean	SE	SD	q2.5	X97.5	n_eff	Rhat
β_0	-0.30	0.00	0.14	-0.58	-0.03	4118	1.00
β_1	0.53	0.00	0.16	0.21	0.84	1122	1.0
σ_u	0.48	0.00	0.09	0.32	0.66	826	1.00
ρ	0.86	0.01	0.17	0.40	1.00	528	1.00
$\mu[5]$	13.24	0.04	2.93	8.40	19.70	6056	1.00
$\phi^Y[5]$	1.35	0.01	0.42	0.55	2.23	1257	1.00
$\theta[5]$	0.24	0.01	0.95	-1.64	2.09	4297	1.00

values in each covariate (X_1 and X_2). In this way, we got three incomplete datasets, each with two sets of missing regions $i \in M_1$ for covariate X_1, and $i \in M_2$ for covariate X_2, such that with 10% missing proportion in each covariate:

M_1 = {14, 25, 31, 42, 43, 50, 51, 67, 79, 97},
M_2 = {9, 26, 57, 69, 72, 90, 91, 92, 93, 99},

with 15% missing proportion:

M_1 = {31 79 51 14 67 42 50 43 97 25 90 69 57 9 72},
M_2 = {26 7 42 9 83 36 78 81 43 76 15 32 99 97 41},

with 20% missing proportion:

M_1 = {63 13 82 97 91 25 38 21 79 41 47 60 16 6 72 39 31 81 50 34},
M_2 = {4 13 69 25 52 22 89 32 97 87 35 40 30 12 31 88 64 14 71 67}.

The proposed model was then applied to these incomplete

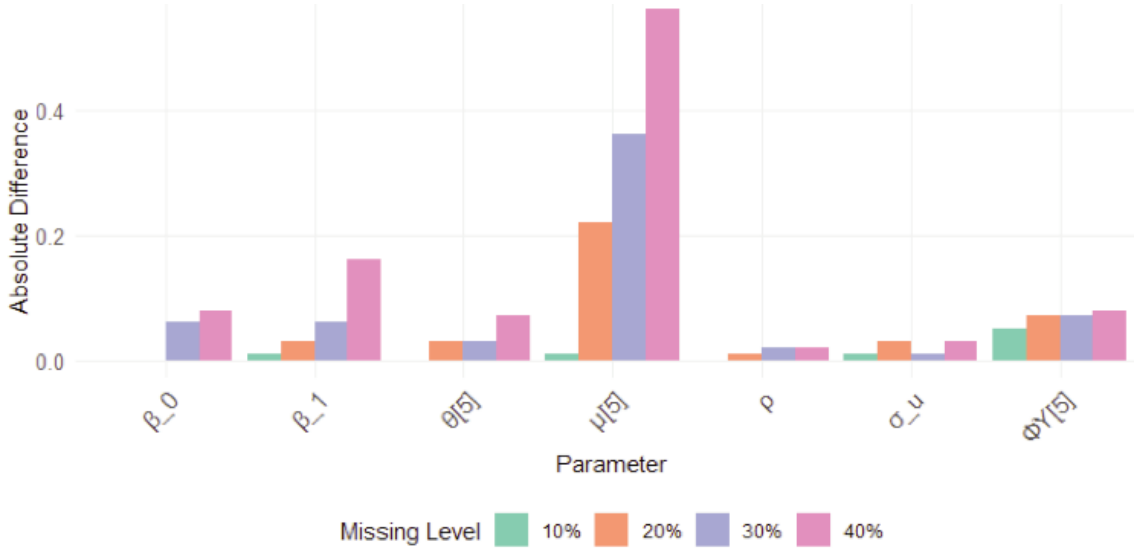

Figure 1. Absolute difference of the proposed model's estimates from the baseline on the Scottish lip cancer dataset.

Table 5. Results of the complete data BYM2 model on the Scottish lip cancer data.

Parameter	Mean	SD	q5	q95	ess_bulk	Rhat
β_0	-0.217	0.129	-0.431	-0.0057	1927	1.00
β_1	0.365	0.134	0.143	0.583	2065	1.00
σ_u	0.514	0.088	0.383	0.675	739	1.00
ρ	0.878	0.144	0.572	1	686	1.00
μ[5]	13.80	3.13	9.14	19.4	5079	1.00
$\phi^{\wedge Y}$[5]	1.43	0.406	0.789	2.14	1255	1.00
θ[5]	0.170	0.595	-1.41	1.75	7680	1.00

Morris, 2019. q5, 5% posterior quantile (lower bound of 90% credible interval); q95, 95% posterior quantile (upper bound of 90% credible interval); ess_bulk, bulk effective sample size (number of independent draws from the posterior).

Table 6. Results at 10% missing proportion in North Carolina SIDS data.

Parameter	Mean	SE	SD	X2.5	X97.5	n_eff	Rhat
β_0	0.01	0.00	0.05	-0.09	0.10	2870	1.00
β_1	-0.05	0.00	0.04	-0.12	0.03	2181	1.00
β_2	0.15	0.00	0.06	0.03	0.25	1484	1.00
σ_u	0.25	0.00	0.07	0.11	0.40	740	1.00
ρ	0.47	0.02	0.34	0.00	1.00	397	1.01
μ[5]	4.55	0.02	1.08	2.68	6.80	1914	1.00
$\phi^{\wedge Y}$[5]	-0.27	0.02	0.67	-1.61	1.02	1106	1.00
θ[5]	-0.27	0.02	0.92	-2.10	1.51	2104	1.00

datasets.

We assigned weakly informative priors based on the fact that all covariates were scaled to have mean 0 and variance 1, which places all parameters on comparable scales. Accordingly, we used priors cantered at zero with small to moderate variances as follows: β_0, β_1, β_2, $_{\sigma u}$~N(0,1); μ_{Xj}~N(X_j^{obs},0.5); σ_{Xj}~N(0,0.5); σ~N(0,0.1); ρ,ρ_j~Beta(0.5,0.5). The model was implemented in 'RStan' using 4 chains, each run for 2,000 iterations, with the first 1000 for 'warm-up', yielding a total of 4, 000 posterior draws. The R codes for the two-covariate model are given in Appendix B in the supplementary file.

The results of the proposed model on these incomplete datasets are shown in Tables 6-8, presenting the posterior summaries of model parameters of interest, as well as a fitted value μ[5].

In addition to the incomplete data analyses, we analyzed the complete SIDS dataset to obtain baseline parameter estimates as given in Table 9. The estimates obtained from incomplete datasets in Tables 6-8 were compared to their respective baseline values to assess the recovery of the parameters of interest.

Figure 2 compares absolute deviations of the parameter estimates from the baseline at different missing proportions. While all parameter estimates are consistent at 10% and 15%, the area-specific mean μ[5] deviated sharply at 20% vacancy in each covariate, followed by σ_u.The posterior SD and the width of the CIs in Tables 6-8 indicate that the uncertainty remained stable for all parameters of interest across varying missing proportions. For all parameters, *Rhat* values were ≤1.01, and the effective sample size exceeded 300 for all missing proportions. Importantly, unlike the Scottish dataset, the North Carolina SIDS results produced no diagnostic warnings (e.g., divergent transitions, low Effective Sample Size (ESS)), indicating convergence for reliable posterior inference. These results suggest that, given the proposed method's large number of parameters in the two-covariate specification, stable convergence requires datasets with 100 or more spatial units

Furthermore, we compared the predictive performance across varying missing proportions, using the Root Mean Square Error (RMSE), defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y})^2}{N}} \quad \text{Eq. 11}$$

Table 7. Results at 15% missing proportion in North Carolina SIDS data.

Parameter	Mean	SE	SD	X2.5	X97.5	n_eff	Rhat
β_0	0.01	0.00	0.05	-0.09	0.11	2408	1.00
β_1	-0.06	0.00	0.04	-0.13	0.02	833	1.00
β_2	0.13	0.00	0.05	0.03	0.24	799	1.01
σ_u	0.25	0.00	0.07	0.10	0.40	507	1.01
ρ	0.43	0.02	0.34	0.00	1.00	365	1.01
μ[5]	4.45	0.02	1.05	2.66	6.77	2139	1.00
ϕ^Y[5]	-0.22	0.03	0.67	-1.51	1.17	602	1.00
θ[5]	-0.21	0.02	0.91	-2.07	1.58	2909	1.00

Table 8. Results at 20% missing proportion in North Carolina SIDS data.

Parameter	Mean	SE	SD	X2.5	X97.5	n_eff	Rhat
β_0	0.01	0.00	0.05	-0.09	0.10	2737	1.00
β_1	-0.06	0.00	0.04	-0.13	0.02	2106	1.00
β_2	0.17	0.00	0.05	0.07	0.27	2095	1.00
σ_u	0.19	0.00	0.08	0.02	0.35	691	1.01
ρ	0.41	0.01	0.35	0.00	1.00	838	1.00
μ[5]	4.99	0.02	1.05	3.02	7.07	2399	1.00
ϕ^Y[5]	-0.27	0.03	0.70	-1.63	1.11	750	1.01
θ[5]	-0.25	0.01	0.95	-2.07	1.70	4650	1.00

Table 9. Results of the complete data BYM2 model on the North Carolina SIDS dataset.

Parameter	Mean	SE	SD	X2.5	X97.5	n_eff	Rhat
β_0	0.02	0.00	0.05	-0.08	0.12	2325	1.00
β_1	-0.06	0.00	0.04	-0.12	0.02	2478	1.00
β_2	0.13	0.00	0.05	0.03	0.23	2547	1.00
σ_u	0.25	0.00	0.07	0.12	0.40	995	1.00
ρ	0.40	0.02	0.33	0.00	0.99	367	1.00
μ[5]	4.43	0.02	1.04	2.60	6.66	4102	1.00
ϕ^Y[5]	-0.27	0.02	0.66	-1.59	1.06	925	1.00
θ[5]	-0.25	0.01	0.93	-2.07	1.53	5477	1.00

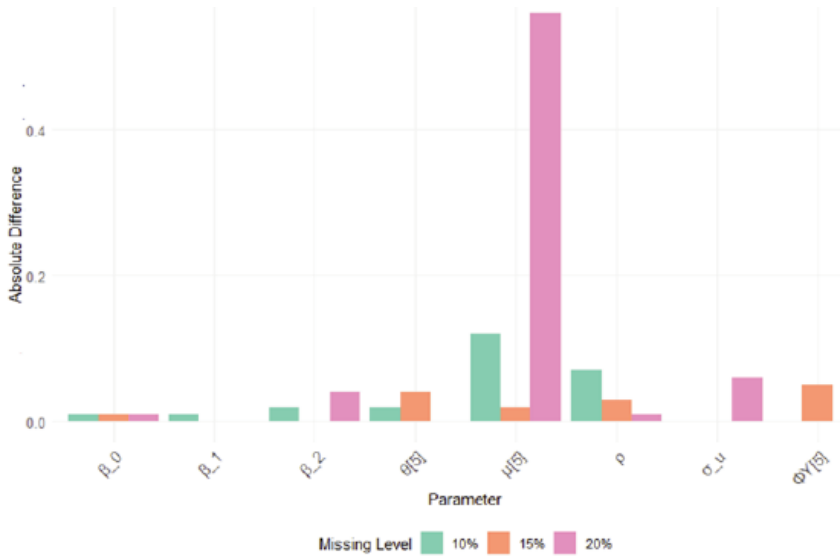

Figure 2. Absolute difference of the proposed model's estimates from the baseline on the North Carolina SIDS dataset.

where *Yi* represents the observed SIDS counts and $\hat{Y}$ denotes the predicted value for *i*-th spatial unit.

The RMSE values were compared with the baseline value (2.229) of the complete-data Bym2 model on the North Carolina SIDS dataset. The proposed model achieved RMSE values of 2.231, 2.230, and 2.385 at 10%, 15%, and 20% missing proportions per covariate, respectively. These results demonstrate that the proposed BYM2 model maintains robust predictive performance up to 15% missing data, with a sharp decline at 20% missing data per covariate.

These results suggest that in scenarios with missing values in two covariates, the proposed model can recover parameters of interest with a missing proportion of up to 15% in each covariate (equivalently, a cumulative 30% across both covariates). Consistent with our previous analyses of missing values in a single covariate, which showed good recovery with a missing proportion of up to 30%, the current results demonstrate effective recovery of parameter estimates in scenarios involving multiple partially observed covariates with a cumulative missing proportion of up to 30% across them.

Conclusions

We propose a Bayesian spatial model for disease mapping in the presence of missing covariates. The proposed model extends the prominent BYM modelling framework to accommodate the missing covariate values while modelling disease counts. Evaluation at the increasing levels of missing data indicates that the proposed model performs well with a missing proportion of up to 30% in a single covariate or a cumulative 30% across multiple partially observed covariates. The proposed method performs reliably for spatial datasets of adequate size, whereas smaller datasets may lead to convergence issues such as divergent transitions and low effective sample sizes.

References

Besag J, York J, Mollié A, 1991. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 43:1–20.

Best N, Richardson S, Thomson A, 2005. A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res 14:35–59.

Bhakkan-Mambir B, Luce D, Deloumeaux J, 2025. Cancer incidence in the vicinity of open landfills in Guadeloupe, French West Indies. BMC Public Health 25:1–8.

Cifo D, Estévez-Reboredo RM, González-Barrio D, Jado I, Gómez-Barroso D, 2024. Epidemiology of Q fever in humans in four selected regions, Spain, 2016 to 2022. Eurosurveillance 29:2300688.

Clayton D, Kaldor J, 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43:671-81.

Cressie N, Chan NH, 1989. Spatial modeling of regional variables. J Am Stat Assoc 84:393–401.

Dean CB, Ugarte MD, Militino AF, 2001. Detecting interaction between random region and fixed age effects in disease mapping. Biometrics 57:197–202.

Haque S, Price A, Mengersen K, Hu W, 2025. Evaluating the impact of the modifiable areal unit problem on ecological model inference: a case study of COVID-19 data in Queensland, Australia. Infect Dis Model 10:1002-19.

Lee D, 2011. A comparison of conditional autoregressive models used in Bayesian disease mapping. Spat Spatiotemporal Epidemiol 2:79–89.

Leroux BG, Lei X, Breslow N, 2000. Estimation of disease rates in small areas: a new mixed model for spatial dependence. Statistical models in epidemiology, the environment, and clinical trials, New York, pp. 179-191.

Leveau CM, Riancho J, Shaman J, Santurtún A, 2024. Spatial analysis of ischemic stroke in Spain: the roles of accessibility to healthcare and economic development. Cad Saude Publica 40:e00212923.

Ma Z, Hu G, Chen MH, 2021. Bayesian hierarchical spatial regression models for spatial data in the presence of missing covariates with applications. Appl Stoch Models Bus Ind 37:342–59.
MacNab YC, 2011. On Gaussian Markov random fields and Bayesian disease mapping. Stat Methods Med Res 20:49–68.
Morris M, 2019. spatial models in stan: intrinsic auto-regressive models for areal data. Stan case study. Accessed 1 January 2026. Available from: https://mc-stan.org/learn-stan/case-studies/icar_stan.html
Morris M, Wheeler-Martin K, Simpson D, Mooney SJ, Gelman A, DiMaggio C, 2019. Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan. Spat Spatiotemporal Epidemiol 31:100301.
Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH, 2017. Penalising model component complexity: A principled, practical approach to constructing priors. Statist Sci 32: 1-28
Slater JJ, Brown PE, Rosenthal JS, Mateu J, 2022. Capturing spatial dependence of COVID-19 case counts with cellphone mobility data. Spat Stat 49:100540.
Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC, 2021. Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. Bayesian analysis 16:667-718
Wakefield J, 2007. Disease mapping and spatial regression with count data. Biostatistics 8:158–83.

Online supplementary materials
Appendix A: R codes for analyzing the Scottish lip cancer dataset (single covariate).
Appendix B: R codes for analyzing the North Carolina SIDS data set (two covariates).

Received: 27 March 2026; Accepted: 16 April 2026.

Contributions: the authors contributed equally to the present work.

Conflict of interests: the authors declare that there is no conflict of interest.

Ethical approval: not applicable

Availability of data and material: the Scottish lip cancer data used in this study are available from the published article(Clayton and Kaldor, 1987) . The data were sourced from the scotland_data.R script provided in the Stan example models repository (Stan Development Team) available at: https://github.com/stan-dev/example-models/blob/master/knitr/car-iar-poisson/scotland_data.R). The North Carolina SIDS dataset available in the published article (Cressie and Chan, 1989) can be accessed via the `sf` R package (Pebesma, 2018) using the following command:
nc <- sf::st_read(system.file ("shape/nc.shp", package, "sf"))

Funding: this work was supported by the National Social Science Fund of China under grant No.21BTJ041.