Scientific African 27 (2025) e02498

Contents lists available at ScienceDirect

Scientific African

journal homepage: www.elsevier.com/locate/sciaf

Evaluating the impact of misspecified spatial neighboring structures in Bayesian CAR models

Ernest Somua-Wiafe [a], Richard Minkah [a], Kwabena Doku-Amponsah [a], Louis Asiedu [a], Edward Acheampong [a], Samuel Iddi [a,b,*]

[a] University of Ghana, Department of Statistics and Actuarial Science, Accra, Ghana
[b] African Population and Health Research Center, APHRC Campus, Manga Close, Off Kirawa Road, P.O. Box 10787-00100, Nairobi, Kenya

ARTICLE INFO

Editor name: Hilary Izuchukwu Okagbue

Keywords:
Conditional autoregressive
Model misspecification
Moral index
Random effects
Road crash
Spatial modeling

ABSTRACT

Spatial neighboring graphs play a crucial role in accounting for global spatial dependency, particularly in spatial models that utilize the Conditional Autoregressive (CAR) covariance structure. The Bayesian modified Besag–York–Mollié (BYM2) model, which falls under the category of CAR models, introduces a precision parameter to quantify the variability not captured by the fixed risk components and a mixing parameter to decipher the proportion of random effects attributed to the spatial component and the aspatial random noise. Despite the advantages these extra features bring, misspecification of BYM2 model components is common, and its effects are not well understood. Previous studies often avoid simulations due to computational demands, relying instead on performance metrics for inferences and model comparisons using empirical data.

This study uses comprehensive simulations to examine the impact of erroneously specified spatial neighborhood structures on the BYM2 model. We considered three different neighborhood structures: a first-order adjacency-based structure and two minimum distance-based structures with threshold distances of 70 km and 140 km at various sparsity levels. For each structure, we simulate data under that structure and then analyze it using the remaining two structures as misspecified cases to evaluate their impact on model fit. Fixed PC prior settings were applied to control for prior specification effects in examining bias and MSE. The study was further validated through practical analyses of road crash incidents in Ghana and a lip cancer cases data in Scotland, UK.

Our findings reveal that incorrect specification of the neighboring structure does not significantly impact the fixed effects. However, it affects the estimates of the mixing parameter and precision term, thus impacting the spatial component. In cases of high spatial dependency and misspecified neighborhood structures, the BYM2 model tends to underestimate the mixing parameter. Under-specifying the neighborhood structure results in underestimated hyper-parameter values while over-specifying it leads to an overfitted spatial smooth. The empirical application results which were consistent with the simulation also emphasized the critical importance of accurately specifying spatial structures in BYM2 models. Relying solely on metrics like the Watanabe-Akaike Information Criterion (WAIC), Deviance Information Criterion

* Corresponding author at: African Population and Health Research Center, APHRC Campus, Manga Close, Off Kirawa Road, P.O. Box 10787-00100, Nairobi, Kenya.

E-mail addresses: esomua-wiafe@ug.edu.gh, ernestsoms@gmail.com (E. Somua-Wiafe), rminkah@ug.edu.gh (R. Minkah), kdoku-amponsah@ug.edu.gh (K. Doku-Amponsah), lasiedu@ug.edu.gh (L. Asiedu), eoacheampong@ug.edu.gh (E. Acheampong), siddi@ug.edu.gh, siddi@aphrc.org (S. Iddi).

https://doi.org/10.1016/j.sciaf.2024.e02498
Received 31 October 2023; Received in revised form 7 November 2024; Accepted 9 December 2024

Available online 19 December 2024

(DIC), and Conditional Predictive Ordinate (CPO) estimates to determine an optimal spatial structure can be misleading. Instead, the Moran's Index (MI) statistic is more reliable for identifying the most suitable neighborhood structure.

Introduction

In recent years, spatial modeling techniques have become increasingly popular across a variety of disciplines including environmental science, health science, economics, and ecological studies among others. Since events occurring in space and time may be interrelated, it is crucial that the developed model accurately captures the amount of spatial variability exhibited within the random process. In the literature on spatial modeling, there are two main classes of spatial areal models which utilize spatial neighboring structures and they include the Simultaneous Autoregressive (Spatial Autoregressive (SAR)) models and the Conditional Autoregressive (CAR) models. The SAR models are widely accepted in econometrics and are mostly used to fit continuous responses whereas the CAR models are commonly accepted in Epidemiology and are frequently used to model discrete responses.

The choice of spatial weighting in spatial areal modeling lacks a universal principle and is often based on researcher judgment or computational convenience. Some analysts also consider a measure of spatial autocorrelation, while others rely on model assessment metrics. Hence, analysts inevitably face the risk of misspecifying the true spatial structure while fitting their spatial areal model. Although literature extensively covers misspecification in SAR models, there have been limited studies on misspecification when utilizing CAR models. The Bayesian Spatial CAR models typically involve complex calculations, making it computationally demanding to explore the effects of misspecification across a wide range of scenarios through simulations. As a result, researchers are often constrained to take an empirical route, heavily relying on performance metrics such as Deviance Information Criterion (DIC) and Watanabe–Akaike Information Criterion (WAIC) for model selection and inference.

Earnest et al. [1] conducted a study that involved the creation of twelve spatial weight matrices, which were used to assess their impact on the CAR model when applied to empirical spatial and spatiotemporal data. The study found that spatial smoothing was greatly influenced by the weighting structure considered, with different structures producing contrasting effects. In terms of prediction, the study determined that distance-based weights were dominant compared to other weighting structures. Furthermore, the use of covariate-based weights resulted in a superior smoothing effect that identified approximately 74% of the high-risk areas compared to the other specified connectivity matrices. Earnest et al. [1] recommended that the choice of the weighting structure should be guided by a comprehensive study that considers the associated risks and costs, as well as the focus of research.

Likewise, Duncan, White and Mengersen [2] utilized seventeen weighting options in fitting the BYM model to both a simulated dataset and an empirical dataset. Their inferences were based on DIC and Moran's Index (MI) values, which consistently showed that the model incorporating the first-order adjacency weight was superior to other BYM models with different weighting options. The authors stressed the importance of accounting for spatial autocorrelation within the model framework and suggested that sparse weights might be adequate for defining credible amounts of spatial interactions in the data. They also highlighted the first-order adjacency weight matrix as a valid option.

Assuncao and Krainski [3] investigated the use of adjacency-based matrices to capture spatial dependency in Bayesian modeling, specifically when using the CAR or Intrinsic Conditional Autoregressive (ICAR) prior density for the spatial component. They decomposed the prior and posterior covariance structures into infinite sums of terms by limiting the correlation parameter's value. They concluded that concerns about the non-intuitive nature of the CAR and ICAR structures are reduced, particularly when the likelihood's precision is higher than that of the prior density. Utilizing regular and irregular areal-referenced connectivity maps, they also discovered that the influence of areas that are farther apart depends significantly on the convergence speed of the infinite sum, which is dependent on the sparsity of the neighborhood matrix, the associated eigenvectors, and the value of the second largest eigenvalues.

Briz-Redon et al. [4] evaluated various spatial structures for capturing the spatial dependency of COVID-19 relative risk across small geographic regions in Valencia, Spain. Using a spatiotemporal BYM model, they compared contiguity-based, distance-based, covariate-based (incorporating movements and sociodemographic features), and hybrid neighborhood structures. They assessed the model's goodness of fit, predictive accuracy, detection of high-risk areas, spatiotemporal autocorrelation, and smoothing capabilities. Results indicated that the contiguity-based structure, along with some distance-based and sociodemographic structures, outperformed those based on k-nearest neighbors and mobility flows. Additionally, combining or adjusting neighborhood structures did not improve model performance.

Regarding SAR misspecification, Betz et al. [5] compared the biases that arise when one fails to account for spatial dependency against using the Spatial Lag Model (SLM) with misspecified spatial weights and established that it is advantageous to opt for the model with misspecified spatial weights over the non-spatial model even in the worse case of spatial misrepresentation. They therefore added that the gains realized over the non-spatial model seem to increase with a decreasing likelihood of spatial weight misspecification. Thus with limited information about the spatial structure, Betz et al. [5] encouraged analysts to pursue spatial analysis rather than taking the non-spatial route.

Juhl [6] likewise investigated the issue of spatial structure uncertainty in spatial econometric models, where specifying the spatial dependency matrix $\mathbf{W}$ is often challenging due to limited theoretical guidance. Using Monte Carlo experiments, they demonstrated that while effect estimates remain robust to misspecification in $\mathbf{W}$'s structure, inaccuracies in defining neighborhoods can introduce

bias in SAR models. Therefore, they proposed Bayesian Model Averaging (BMA) as a solution. The study observed that BMA effectively addresses such uncertainties and identifies the true spatial structure.

Other authors have focused on covariate misspecification in recent years. Amongst such studies, Arambepola et al. [7] examined the impact of model misspecification as a result of unobserved covariates in the context of disaggregation in spatial disease mapping. Using simulated data under various conditions, the performance was assessed across different sample sizes, aggregation levels, and degrees of model misspecification. Results showed that well-specified models provided accurate fine-scale predictions even with fewer observations and larger areas. However, under model misspecification, predictive performance declined for larger areas but remained robust for smaller regions.

Previous studies on spatial model misspecification utilizing extensive simulations have focused on SAR models and other formulations, either by using erroneous spatial weights [5,6,8–10] or by considering the omission of spatial covariates [7] and addressing these issues through a spatial random component. Those focusing on the Bayesian CAR structure often compare the effects of several spatial structures using goodness-of-fit measures to determine the most suitable spatial structure for the empirical data [1,2,11]. This study aims to fill that gap by conducting a simulation study on the BYM2 model, which features a unique CAR structure, to ascertain the influence of the choice of spatial neighborhood structure on model performance and inference. Additionally, it seeks to identify mechanisms that could enhance the researcher's ability to select the most appropriate spatial structure for the phenomenon under investigation.

The BYM2-CAR model, initially proposed by [12], is a common model that incorporates a scaled spatial component for model comparison across different neighborhood structure structures and reduces the confounding effect between the precision and mixing parameters, which are components of random effects. Because of its strengths, we have seen its applications in several studies [13,14] and its extensions [15,16] yet there have been limited studies that explore the consequences of misspecifying its components. This study therefore employs a simulation approach to investigate the performance of the BYM2 model in the presence of a misspecified spatial network when analyzing Poisson responses aggregated over areal support. We limited the spatial structures to contiguity-based and distance-based structures and utilized the Integrated Nested Laplace Approximation (INLA) to control for model runtime.

The impact of misspecification of spatial models is an active area of research and therefore findings of this study are expected to bolster practitioners' confidence in spatial modeling techniques, which are relatively newer than non-spatial approaches that fail to capture the additional variability resulting from spatial influence. Additionally, this study will serve as a valuable reference for researchers interested in more advanced spatial modeling techniques, which necessitate the careful selection of neighborhood structures. It also highlights areas where further investigations are needed to develop appropriate testing procedures. Moreover, this study aims to fill the gap in the literature concerning the impact of misleading neighborhood structures on modeling spatial outcomes and model selection, particularly in the context of count responses.

This paper is structured in the following manner: In Section "Methodology", we discuss the BYM2 model framework and explore the adjacency-based and minimum-distance-based neighborhood structures that represent the spatial networks utilized in the study, as well as the need to scale the spatial component. The latter part of Section "Methodology" briefly discusses theoretically the nature of errors introduced in the BYM2 model in the presence of misspecified spatial weights and also reviews the key metrics used for model assessment. In Section "Simulation Results and Applications", we investigate the potential consequences of misspecifying the neighborhood structure in the presence of scaling, using pseudo-data to illustrate the intuitive implications. Additionally, Section "Simulation Results and Applications" uses different spatial neighborhood structures to analyze road crash incidents in Ghana and the lip cancer study in Scotland, UK, and compares the resulting model outputs. Finally, the paper concludes in Section "Conclusion" and suggests areas for future research.

Methodology

The BYM2 model

Simpson et al. [12] proposed a modification to the BYM model Besag, York and Mollie [17], incorporating a spatial control parameter and a scaled spatial effect $\boldsymbol{u}_*$ that scales the marginal variance to 1, regardless of the neighborhood structure utilized, thus addressing scaling and identifiability issues regarding the structured and unstructured random effects. Specifically, the mixing parameter $\phi \in [0,1]$ was introduced into their model framework to balance the level of variability attributed to the spatially structured effect and the non-spatial random effect.

Assume that the count responses $y_i, i = 1,2,\ldots,n$ are aggregated at the areal level and generated from a Poisson distribution with mean μ_i, the BYM2 spatial count model can be mathematically represented as:

$$log(\mu_i) = \beta_0 + \mathbf{z}_i^T\boldsymbol{\beta} + \tau_\epsilon^{-0.5}\left[(1-\phi)^{0.5}\boldsymbol{v} + \phi^{0.5}\boldsymbol{u}_*\right] \tag{1}$$

where β_0 is the intercept, $\boldsymbol{\beta}$ represents the vector of slope coefficients for the area-specific covariates $\boldsymbol{z}_i$, $\phi \in [0,1]$ denotes the spatial control parameter, τ_ϵ represents the precision parameter for the combined random effect $\boldsymbol{\epsilon}$, $\boldsymbol{u}_*$ represents the scaled structured component, which models the positive spatial interactions specified as $\boldsymbol{u}_* \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_*^-)$ and $\boldsymbol{v}$ represents a vector of spatially unstructured random effects modeled as $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

In this context, $\mathbf{Q}_*^-$ refers to the generalized inverse of the scaled precision matrix $\mathbf{Q}_*$, which embodies the specified spatial network. The unscaled $\mathbf{Q}$, which equals $(\mathbf{D} - \mathbf{W})$ where $\mathbf{W}$ represents an $(n \times n)$ connectivity matrix, and $\mathbf{D}$ is an $(n \times n)$ diagonal matrix with entries as the number of neighbors of region i, can be, in other words, defined as:

$$Q_{ij} = \begin{cases} |N(i)| & \text{if } i = j \\ -1 & \text{if } i \sim j \ \text{ (ie. region } \ i \ \text{ and } \ j \ \text{ are neighbors)} \\ 0 & \text{otherwise} \end{cases}$$

where $|N(i)|$ denotes the number of neighbors area i has.

For the purpose of this study, we specify two classes of connectivity matrices $\mathbf{W}$: Adjacency-based connectivity and Minimum distance-based connectivity.

Two well-known adjacency-based structures used in representing areal spatial networks are the Rook's and Queen's structures. The Rook's structure considers two regions, s_i and s_j, as neighbors if they share at least one common spatial edge, whereas the Queen's structure considers them neighbors if they share at least one common spatial vertex (point). In both cases, a weight of 1 is assigned to w_{ij} if s_i and s_j are neighbors and 0 otherwise [1,2,18–20].

In contrast, distance-based structures aim to define neighbors based on a function of geographical distance. When working with area-referenced data, mostly, the Euclidean distance between two areal units is measured as the interval between their centroids. The minimum distance structure is defined by a threshold distance, denoted as d (which may be fixed or varying), where sets of spatial units within the distance $(0, d]$ are considered neighbors, while those outside the radius are not. The matrix $\mathbf{W}$ is constructed with binary entries for n areal units, where $w_{ij} = 1$ if $0 < |s_i - s_j| \leq d$ and $w_{ij} = 0$ if $|s_i - s_j| > d$, with $|s_i - s_j|$ representing the Euclidean distance between spatial units s_i and s_j. Increasing the threshold distance will result in more neighbors and a denser connectivity matrix. Therefore, we explore different values of d while maintaining some level of sparsity to identify a suitable value for our research purposes [21].

The ability to scale $\mathbf{Q}$, which embodies the connectivity matrix $\mathbf{W}$, provides a degree of control over the selection of hyper-prior precision, allowing for more feasible comparisons between models with differing neighborhood structures. If we treat $[\mathbf{Q}^-]_{ii}$ as a diagonal matrix with the same entries as $\mathbf{Q}^-$ on the main diagonal, we can compute the Generalized Variance (GV) based on the marginal geometric mean as follows:

$$\sigma^2_{\text{GV}}(\boldsymbol{\epsilon}) = \phi\tau_\epsilon^{-1} \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log\left([\mathbf{Q}^-]_{ii}\right)\right) \tag{2}$$

This expression is dependent on $[\mathbf{Q}^-]_{ii}$. Therefore, the GV for different models would differ based on the number of neighbors for each region i, irrespective of the value of ϕ, suggesting that the effect of hyper-priors on the spatial smooth may vary across models with different neighborhood structure structures, making the models not directly comparable at that level.

The value of τ_ϵ is crucial in determining the degree of spatial smoothness associated with the model. If τ_ϵ is underestimated, spatial effects may be overfitted due to increased spatial variability. Conversely, if τ_ϵ is overestimated, the spatial smoothness may be less significant and therefore underestimated [22]. To ensure that the hyper-prior's influence on spatial smoothness is consistent across varying neighborhood structures, we set $\sigma^2_{\text{GV}}(\boldsymbol{\epsilon}) = \phi\tau_\epsilon^{-1}$ by scaling the various spatial structures accordingly.

Scaling the spatial field involves inverting a singular matrix $\mathbf{Q}$, assumed to be semi-positive definite. This can be achieved using generalized inverse or sparse matrix algebra. The former method typically requires a threshold to detect zero singular values. The rank deficiency of $\mathbf{Q}$ determines the set of non-zero eigenvalues used to obtain the generalized determinant and inverse $\mathbf{Q}^-$. Alternatively, for the latter option, a small value δ is added to the leading diagonal of $\mathbf{Q}$ to regularize it, and the diagonal terms of $\mathbf{Q}^-$ can be obtained using linear constraints. The resulting diagonal matrix provides the variance terms, serving as a scaling factor for the respective $\mathbf{Q}$.

From the BYM2 formulation, the conditional distribution of the random effects $\boldsymbol{\epsilon}$ may be defined as follows:

$$\boldsymbol{\epsilon} \mid \tau_\epsilon, \phi \sim \mathcal{N}\left(0,\ \tau_\epsilon^{-1}\left[(1-\phi)\mathbf{I} + \phi\mathbf{Q}_*^-\right]\right). \tag{3}$$

Here, the symbol $\mathbf{I}$ represents an identity matrix, τ_ϵ is the marginal precision, and $0 \leq \phi \leq 1$ is the mixing parameter that regulates the level of variability due to the spatially structured effect and that associated with the non-spatial random effect.

According to Rue, Martino and Chopin [23], utilizing a sparse covariance matrix enhances computational efficiency during model implementation. However, the variance–covariance matrix of the random effect incorporates the generalized inverse $\mathbf{Q}_*^-$ of the scaled matrix, which is not necessarily sparse. To maintain the sparsity of the variance–covariance matrix, which provides desirable computational properties, Riebler et al. [22] suggested an augmented parameterization approach for implementation. They posited that the joint distribution of $\boldsymbol{w} = \left(\boldsymbol{w}_1^T,\ \boldsymbol{w}_2^T\right)^T$, with $\boldsymbol{w}_1 = \boldsymbol{\epsilon}$ and $\boldsymbol{w}_2 = \boldsymbol{u}_*$, follows a multivariate normal distribution centered at a mean vector of zeros and a precision matrix given by:

$$\begin{pmatrix} \frac{\tau_\epsilon}{1-\phi}\mathbf{I} & -\frac{(\phi\tau_\epsilon)^{0.5}}{1-\phi}\mathbf{I} \\ -\frac{(\phi\tau_\epsilon)^{0.5}}{1-\phi}\mathbf{I} & \mathbf{Q}_* + \frac{\phi}{1-\phi}\mathbf{I} \end{pmatrix}.$$

This precision matrix preserves the sparsity of $\mathbf{Q}_*$ and aligns with the variance of the distribution of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ after inverting the precision matrix. Consequently, working with the joint $\boldsymbol{w}$ ensures that the sparsity of the spatial structure is retained through $\boldsymbol{w}_2$, enabling a more computationally efficient implementation of the BYM2 model [22].

Misspecification of the BYM2 model as a result of erroneous spatial structure

When referencing a misspecified BYM2 model, we are addressing a situation where the statistical model employed to analyze spatial data aggregated into areal units (such as regions, counties, or districts) fails to effectively capture the inherent spatial

dependencies and heterogeneity within the data. To appropriately account for the spatial variability present in the data, Conditional Autoregressive (CAR) models often incorporate neighborhood structures or spatial weight matrices. In this context, spatial structure misspecification pertains to the erroneous assignment of a scaled precision component $\mathbf{Q}_*^{(MS)}$, which either omits pertinent spatial connections or includes irrelevant ones. Over-specification can occur when the scaled precision matrix $\mathbf{Q}_*^{(OS)}$ contains extraneous connectivity that deviates from the actual structure. Conversely, if essential spatial links are omitted in the fitted scaled matrix $\mathbf{Q}_*^{(US)}$, this is referred to as an under-specification of the true spatial structure.

From the framework of the BYM2 model, we postulate that errors in the neighboring structure $\mathbf{Q}_*^{(MS)}$ are likely to impact the level of spatial variation captured within the model. This, in turn, will lead to a misspecified spatial component $\boldsymbol{u}_*^{(MS)}$ and random effect $\boldsymbol{\epsilon}^{(MS)}$. Consequently, the estimates $\phi^{(MS)}$ and $\tau_\epsilon^{(MS)}$ might also be affected to some extent, depending on the degree of misspecification. However, it is plausible that initially scaling the spatial precision matrix could mitigate the impact on $\tau_\epsilon^{(MS)}$. Additionally, based on the BYM2 formulation, the combined random effect $\boldsymbol{\epsilon}$ is misspecified as $\boldsymbol{\epsilon}^{(MS)}$ due to errors in $\boldsymbol{u}_*^{(MS)}$, which in the long run can affect the estimates of the precision and the mixing parameter. Thus, we can rewrite the combined random component of Eq. (1) as

$$\boldsymbol{\epsilon}^{(MS)} = \left[\tau_\epsilon^{(MS)}\right]^{-0.5}\left(\left(1-\phi^{(MS)}\right)^{0.5}\boldsymbol{v} + \left[\phi^{(MS)}\right]^{0.5}\boldsymbol{u}_*^{(MS)}\right)$$

where $\boldsymbol{u}_*^{(MS)} \sim \mathcal{N}\left(\mathbf{0}, \left[\mathbf{Q}_*^{(MS)}\right]^-\right)$ and, by extension,

$$\boldsymbol{\epsilon}^{(MS)} \mid \tau_\epsilon^{(MS)}, \phi^{(MS)} \sim \mathcal{N}\left(0,\ \left[\tau_\epsilon^{(MS)}\right]^{-1}\left((1-\phi^{(MS)})\mathbf{I} + \phi^{(MS)}\left[\mathbf{Q}_*^{(MS)}\right]^-\right)\right).$$

Given the misspecification of spatial weights, the scaling factor in Eq. (2), which is dependent on $\left[\mathbf{Q}_*^{(MS)}\right]_{ii}^-$, is likely to contain errors of varying degrees, depending on the extent of inaccuracies. Hence it is worth considering whether, to a certain degree, scaling the spatial component can indeed alleviate the challenges posed by misspecification. Moreover, we deduce from Eq. (2) that the GV ie. $\left(\sigma^2_{\mathrm{GV}}(\epsilon)\right)$, which depends on the value of ϕ, could be affected if the misspecification of the spatial network ultimately leads to an underestimation or overestimation of the true value.

Assuming the misspecified spatial structure leads to an underspecified $\boldsymbol{u}_*^{(US)}$, this is likely to result in $\widehat{\phi}^{(US)} \leq \phi$, leading to a spillover into the unstructured component. On the contrary, when the misspecified spatial structure leads to an overly estimated mixing parameter such that $\widehat{\phi}^{(OS)} \geq \phi$, a portion of the unstructured component may be filtered into the structured component, thereby introducing some level of confounding within the random effects. Hence, we expect that in both cases, issues of identifiability would be raised, as the estimate of $\phi^{(MS)}$ would be affected by errors resulting from the nature of neighborhood misspecification. Consequently, this study employs a comprehensive simulation mechanism to investigate the biases and errors introduced by a misspecified neighborhood structure and the extent of damage it inflicts on inference.

Simulation overview

To comprehensively understand the effects a misinformed spatial neighborhood structure poses on model output, we embarked on a simulation exercise.

Based on the 261 districts of Ghana, the study constructs three neighborhood structures with varying levels of connectivity and sparsity, adhering to the symmetric property of the CAR structure. The first structure is based on the Queen's contiguity matrix with a sparsity level of about 2.1%, while the other two neighborhood structures are minimum distance-based contiguity structures, with the second structure being a subset of the third. The second structure was generated by considering neighbors within a radius of 70 km from the centroid of the areal units while ensuring that all areal units had at least one neighbor within its radius. The third neighborhood structure has a wider interval of $(0, 140km]$ resulting in a considerably less sparse matrix.

We selected specific values for the mixing parameter ϕ, including $\phi = 0.30$ representing a lower proportion of spatial dependence and $\phi = 0.90$ indicating a higher proportion. For the precision parameter, we chose $\tau_\epsilon = (4/9,\ 4)$ which are standard deviation values that are ± 0.5 steps away from 1 representing relatively low and high precision. This resulted in four hyper-parameter combinations ($\phi_1 = 0.30$ and $\tau_1 = 4/9$), ($\phi_2 = 0.90$ and $\tau_1 = 4/9$), ($\phi_1 = 0.30$ and $\tau_2 = 4$), and ($\phi_2 = 0.90$ and $\tau_2 = 4$) tested under each of the three neighborhood structures leading to 12 different scenarios. Throughout the model estimation, we maintained the PC prior density concerning our knowledge about the true values of τ and ϕ. We used PC priors that were well-informed about the hyper-parameters being estimated based on our prior knowledge of the true parameter values. For τ_ϵ values of 4/9 and 4, we assigned prior probabilities of $\Pr(\tau_\epsilon^{-0.5} > 1) = 0.9$ and $\Pr(\tau_\epsilon^{-0.5} > 1) = 0.1$, respectively. Additionally, for ϕ values of 0.30 and 0.90, we set prior probabilities of $\Pr(\phi < 0.5) = 0.1$ and $\Pr(\phi < 0.5) = 0.9$, respectively.

Additionally, we generated two covariates and kept their effects fixed throughout the exercise. The first set of covariates consisted of absolute values sampled from the normal distribution $\mathcal{N}(1, 10)$ rounded to 3 decimal places. For the second set of covariates, we generated samples from the log-uniform distribution $log(\mathcal{U}_{[1,3]})$ and rounded them to 3 decimal places. The coefficient values for the fixed effects $\left(\beta_0 = 2,\ \beta_1 = 0.25,\ \text{and}\ \beta_2 = -2\right)$ were kept constant while varying the assigned parameter values for the random components.

For each scenario, we simulated samples based on the BYM2 model structure, noting the true neighborhood structure and parameter values. We then fitted the BYM2 model to the simulated data using all three neighborhood structures, labeling the true structure used for generating the data as the correct structure and the other two as misspecified.

The simulation process was repeated 200 times using the same set of covariates generated for each of the 12 combinations. After several replications of each of the scenarios, we examined model estimates along with other measures such as Bias, MSE, Marginal log-Likelihood (MLIK), WAIC, DIC, and MI statistic.

Coding was done using R version 4.2.3. Model implementation was achieved through the **R – INLA** package and the `inla.scale.model()` function was used to scale the structured component. As a precautionary measure to avoid simulation failures, we chose samples that resulted in a significant spatial dependence and a reasonable set of responses without extreme values.

To promote reproducibility and open science, the scripts used to perform the simulations in this article can be accessed in the GitHub repository: https://github.com/IddiSam/BYM2Simulations.git. The crash data will be provided by the authors upon request.

Metrics for statistical diagnostics and model assessment

Moran's index

It is essential to conduct statistical analysis to determine if there is spatial dependency in the observed data, and Moran's Index [24] is a powerful tool for testing spatial autocorrelation. Spatial autocorrelation between $e(s_i)$ and $e(s_j)$ at locations s_i and s_j measures the similarity of their values, considering their relative positions. The MI statistic I can be computed as:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(e(s_i)-\bar{e}\right)\left(e(s_j)-\bar{e}\right)}{\sum_{i=1}^{n}\left(e(s_i)-\bar{e}\right)^2} \right)$$

where $\bar{e}$ is the mean value, and w_{ij} are the elements of the connectivity matrix. Assuming there is no spatial autocorrelation, the distribution of $\frac{I-E(I)}{\sqrt{Var(I)}}$ is expected to approach a normal distribution with a mean of 0 and a standard deviation of 1, as the sample size n increases. Here, $E(I)$ represents the expected value of the MI statistic, which is equal to $-1/(n-1)$, and $Var(I)$ is the variance of the MI statistic, I, calculated as $E(I^2) - E(I)^2$.

Marginal likelihood

The marginal likelihood, denoted as $p(\boldsymbol{y} \mid \mathbb{M})$, represents the likelihood of observing data for a specific model $\mathbb{M}$. When comparing a set of distinct models that differ in various aspects, such as modeling framework and parameters, we anticipate different values for their marginal likelihoods. Using the INLA, an approximate version of the marginal likelihood, $\tilde{p}(\boldsymbol{y} \mid \mathbb{M}_i)$, can be calculated by integrating an expression involving model parameters and their Gaussian approximation. This approach is effective and accurate [25,26]. Due to the typically small and close-to-zero values of marginal likelihoods, the concept of the MLIK is used. The model with the highest MLIK is considered the most favorable among a group of distinct models, aiding in model selection.

Information criterion (DIC and WAIC)

The Deviance Information Criterion (DIC) is a widely used method for comparing and selecting models. It combines a measure of a model's deviance, denoted as $\mathrm{D}(\boldsymbol{\psi}, \boldsymbol{\vartheta})$, which quantifies how well the model fits the data, with a measure of model complexity. Deviance is defined as $-2\log\left(p(\boldsymbol{y} \mid \boldsymbol{\psi}, \boldsymbol{\vartheta})\right)$, and the effective number of model parameters represents complexity. DIC is calculated as $\mathrm{DIC} = \overline{\mathrm{D}(\boldsymbol{\psi}, \boldsymbol{\vartheta})} + p_D$, where $\overline{\mathrm{D}(\boldsymbol{\psi}, \boldsymbol{\vartheta})}$ is the posterior mean of deviance, and $p_D = \overline{\mathrm{D}(\boldsymbol{\psi}, \boldsymbol{\vartheta})} - \mathrm{D}(\overline{\boldsymbol{\psi}}, \overline{\boldsymbol{\vartheta}})$, where $\overline{\boldsymbol{\psi}}$ and $\overline{\boldsymbol{\vartheta}}$ are the expectation of the model parameters, $\boldsymbol{\psi}$ (vector containing the fixed parameters) and $\boldsymbol{\vartheta}$ (a vector containing the hyper-parameters of random effects), respectively. Additionally, using a point-wise log-likelihood from posterior simulations, the Watanabe–Akaike Information Criterion (WAIC) proposed by Watanabe [27] which is an extension of the well-known Akaike Information Criterion (AIC) is also commonly used. Both WAIC and DIC are considered informative criteria for Bayesian model selection and predictions, as emphasized by Gelman, Hwang and Vehtari [28]. Often, the model with the smallest DIC and WAIC values is preferred.

Estimating the bias and MSE of $\hat{\tau}$ and $\hat{\phi}$

We can evaluate the measure of bias for $\hat{\tau}$ and $\hat{\phi}$ as follows: For τ and ϕ, the bias can be expressed respectively as

$$\mathrm{bias}\left(\hat{\tau}\right) = \frac{1}{R}\sum_{i=1}^{R}\left(\hat{\tau}_i - \tau\right)$$

and

$$\mathrm{bias}\left(\hat{\phi}\right) = \frac{1}{R}\sum_{i=1}^{R}\left(\hat{\phi}_i - \phi\right).$$

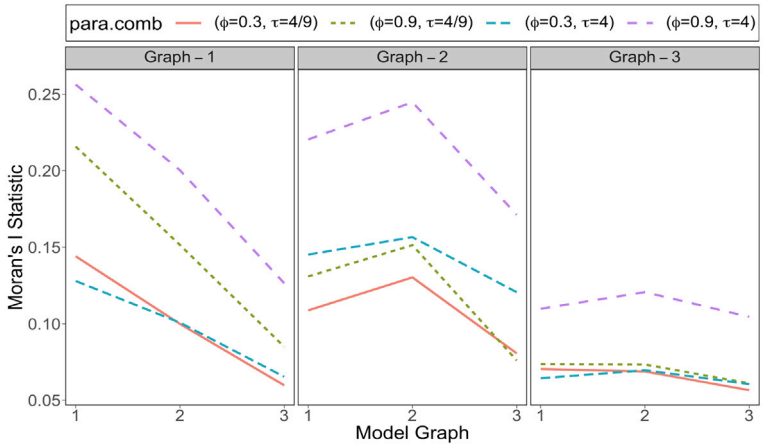
where R represents the number of replications. The MSE of the estimates for $\hat{\tau}$ and $\hat{\phi}$ can be expressed respectively as:

$$\mathrm{MSE}\left(\hat{\tau}\right) = \frac{1}{R-1}\sum_{i=1}^{R}\left(\hat{\tau}_i - \bar{\tau}\right)^2 + \left(\frac{1}{R}\sum_{i=1}^{R}\hat{\tau}_i - \tau\right)^2$$

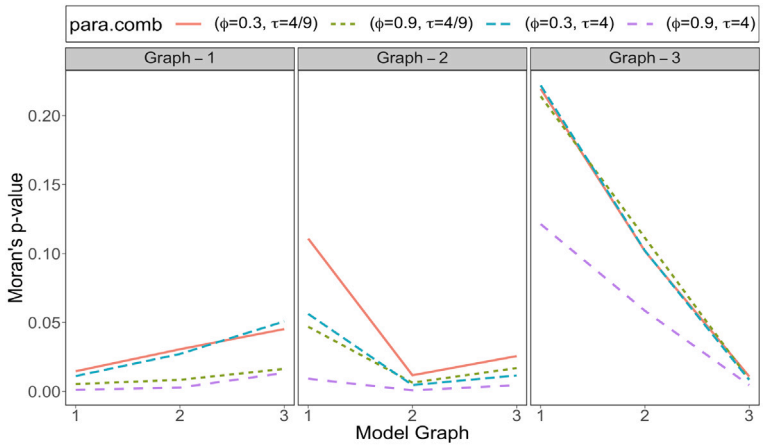
and

$$\mathrm{MSE}\left(\hat{\phi}\right) = \frac{1}{R-1}\sum_{i=1}^{R}\left(\hat{\phi}_i - \bar{\phi}\right)^2 + \left(\frac{1}{R}\sum_{i=1}^{R}\hat{\phi}_i - \phi\right)^2.$$

Here, $\bar{\tau}$ and $\bar{\phi}$ represent the average estimates of τ and ϕ over the R replications.



(a) Moran's I statistic



(b) Moran's p-value

Fig. 1. Plot the average values of Moran's I statistic and p-value for the residuals from the aspatial model, considering the various simulated data generated under the respective spatial networks against the neighboring structure utilized for the test. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

Simulation results and applications

What is the impact of misspecifying the neighborhood structure on the BYM2 model?

This analytic study uses the synthetic data from the BYM2 model over Poisson spatial counts, focusing on the 261 districts of Ghana as the risk surface. After generating the dataset at each stage, we fitted it to a non-spatial Poisson log-linear model, and the residuals were analyzed for spatial autocorrelation using the spatial weights from the three neighborhood structures. The MI measure of spatial dependency was used, and the estimates of MI statistic and p-values were averaged and plotted across the three neighborhood structures. The results indicate that the MI statistic is more effective in identifying the true neighborhood structure when the weight matrix is sparser (see Fig. 1(a)). However, the Moran's p-value was found to be more reliable in identifying the true spatial structure when the structure is less sparse, as shown in Fig. 1(b). Across all scenarios examined, the spatial weights derived from the accurate neighborhood structure yielded the lowest average p-values, implying evidence of spatial autocorrelation as compared to the p-values associated with the incorrect structures. However, this evidence is less pronounced when the actual neighborhood structure is sparser.

To model the simulated dataset, we specified the correct neighborhood structure and compared the results to scenarios where errors existed in the specified spatial structure. The distribution of fixed estimates β_0, β_1, and β_2 were compared using Figs. A.1, A.2, and A.3, respectively. We observed that despite neighborhood structure misspecification, the estimates had similar distributions. Upon inspection, we noted that the distribution of fixed estimates had reduced variability when $\tau = 4$, compared to when τ was approximately 0.444.

Upon examining the estimates of τ, we noted differences in the distribution of model estimates between the correctly specified neighborhood structure and the other two neighborhood structure specifications, especially when $\phi = 0.9$ (refer to Figs. A.4 and

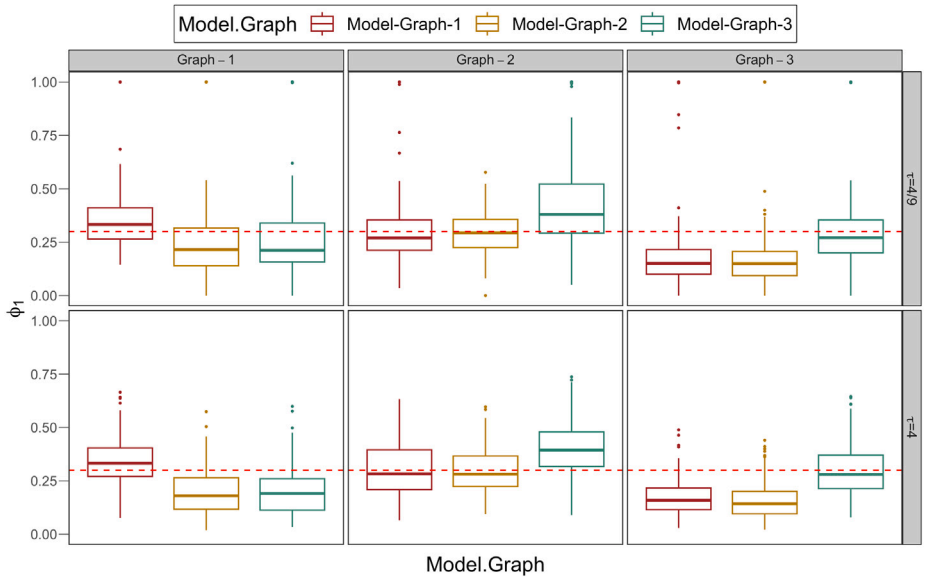


Fig. 2. Distribution of posterior mean of ϕ_1 from simulated results with misspecified neighborhood structures. The horizontal dashed line signifies the true parameter value. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

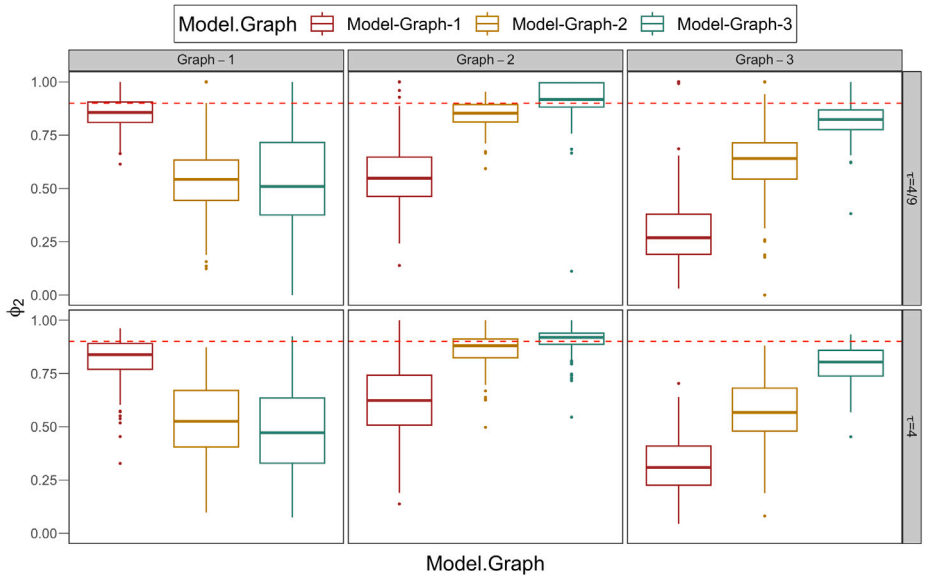


Fig. 3. Distribution of posterior mean of ϕ_2 from simulated results with misspecified neighborhood structures. The horizontal dashed line signifies the true parameter value. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

A.5). This suggests that higher spatial dependency increases the likelihood of a misspecified neighborhood structure introducing errors in estimating the precision value. The plots indicate that an underspecified spatial structure is more likely to result in an underestimation of the precision value, while an over-specified spatial structure is more likely to lead to an overestimated value of the precision. Furthermore, the boxplot for the estimates of τ from the models with the correct spatial weight specifications shows less spread over the range than the distribution of estimates from the misspecified models. This trend is more pronounced when $\phi = 0.9$.

Moreover, we observed that spatial weight misspecification led to biased estimates of ϕ, as illustrated in Figs. 2 and 3. For example, in the dataset generated from Structure 1, specifying the neighborhood structure as Structure 2 or Structure 3 resulted in an underestimation of ϕ. When modeling the data generated from Structure 2, using Structure 1 is more likely to result in

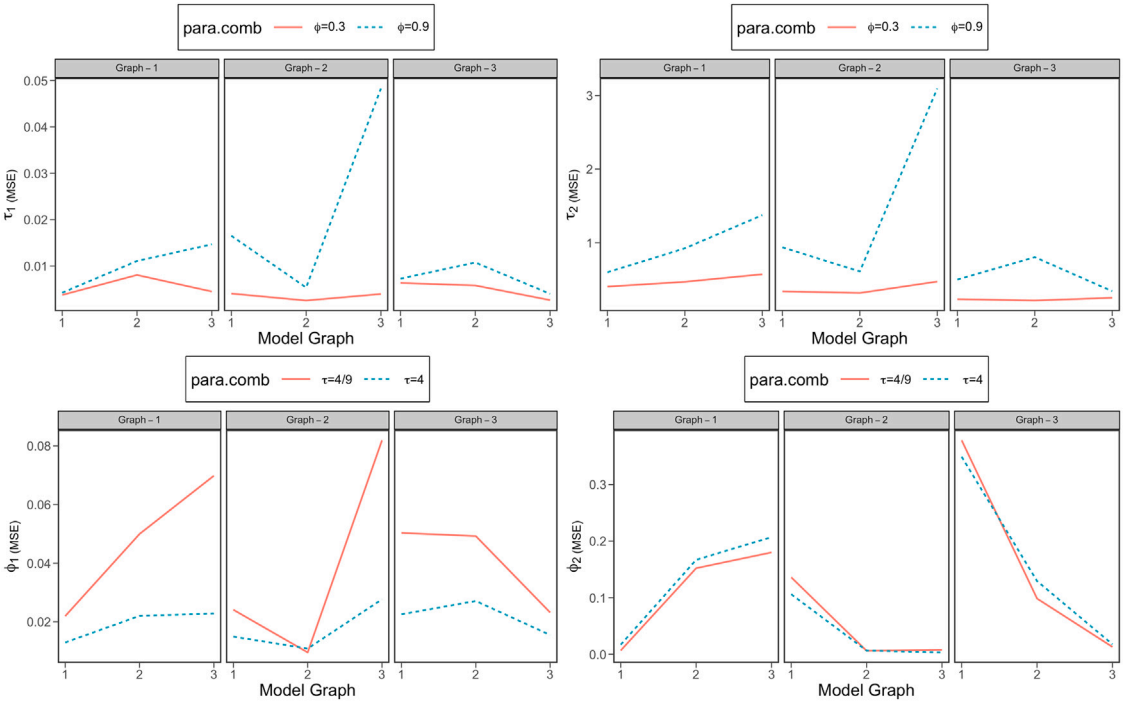


Fig. 4. Plot of MSE estimates of precision and mixing parameter against neighborhood structure misspecification. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

an underestimation of the mixing parameter value, especially when spatial dependence is high. On the other hand, an over-specified Structure 3, which belongs to the same structural class as the true spatial network tends to overestimate the value of ϕ, especially when spatial dependency is low. For the data generated under Structure 3, both Structure 1 and Structure 2 resulted in an underestimated value of the mixing parameter.

After analyzing the bias and MSE plots, it is evident that although the bias estimates do not always support the model with the correct neighborhood structure specification, the MSE estimates consistently favor the model with the correct neighborhood structure specification across all the scenarios studied. The smallest MSE estimate was obtained for the model that employed the true neighborhood structure, as shown in Fig. 4 for each of the scenarios considered. The bias plot of τ_1 supported the correct neighborhood structure specification in all cases except for the dataset generated under Structure 2 with $\phi = 0.3$, where the more sparse Structure 1 was preferred. For the bias estimates of ϕ_1, the correctly specified model graph was endorsed for data generated under Structure 1 with $\tau = 4$ and Structure 3, but the others pointed to one of the misspecified models. Similarly, the bias estimates of ϕ_2 supported the model with the correct neighborhood structure specification for all datasets except for the one generated under Structure 2, which preferred the overly specified Structure 3. In situations where the bias failed to endorse the correctly specified model, we found that the difference in estimates for the chosen specification compared to the corresponding bias estimates from the truly specified model was insignificant. Analyzing the bias estimates for data generated under spatial structure 2, we observe that the estimates of τ and ϕ consistently appear biased upwards when the spatial network is over-specified using Structure 3. On the other hand, based on the data generated under Structure 3, we realize that the estimates of τ and ϕ are negatively biased when the spatial network is under-specified using Structure 2, endorsing our initial claim.

Focusing now on Figs. A.6, A.7, A.8, and A.9, which show the WAIC, DIC, CPO, and MLIK plots, respectively, we note that all the different neighborhood structure specifications exhibit very similar values. Therefore, while these measures may have some role in model selection depending on the research question, they do not possess the capability to distinguish the correct spatial structure model and could potentially be misleading.

Analysis of road crashes in Ghana

As part of our investigation, we analyzed road crash data from Ghana that was collected over 16 years from 2005 to 2020. This data includes all vehicular crashes and pedestrian knockdowns that occurred at the district level and was obtained from the Motor Traffic and Transport Department (MTTD) of the Ghana Police Service by the Building and Road Research Institute (BRRI).

The perception generally is that population size could play a key role in explaining crash frequency. It is generally expected that higher population densities lead to increased vehicle and pedestrian traffic, especially in urban districts, which in turn raises the likelihood of vehicular crashes. In Ghana, most districts are rural with few motorable road networks and are largely forested

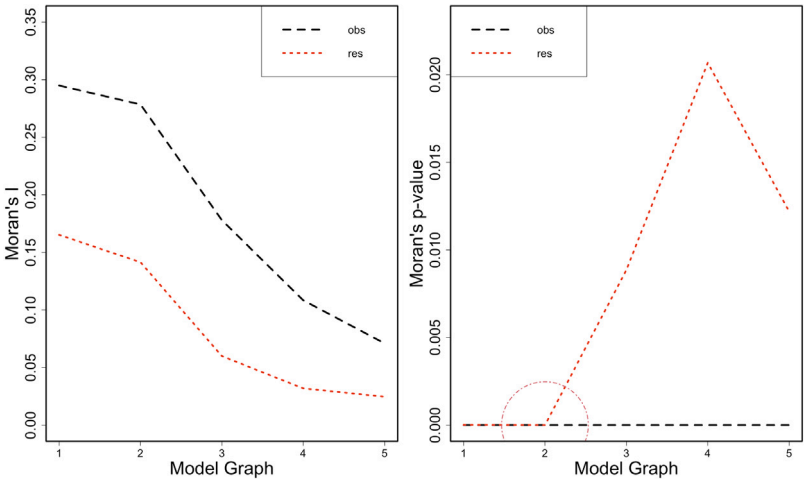


Fig. 5. Moran's I statistic (left) and p-value (right) for crash counts and the residuals from the non-spatial model.

and expansive. It would be interesting to explore how district size affects road crash occurrences. Additionally, the presence of police stations might indicate the level of law enforcement and traffic regulation, which could be crucial in understanding crash frequency. Some districts shared a common reporting station, so they were merged, resulting in a total of 255 areal units for analysis. Our explanatory variables include the district's population size and area (in km^2), as well as the number of MTTD stations at the district level. The response variable in this study is the total number of road crashes, which is depicted in Fig. A.10 as being heavily skewed to the right at the district level in Ghana. Furthermore, the mean value of approximately 756 crashes and a variance that is about 2378 times larger than the mean suggest a potential issue of over-dispersion.

Here, we directly model the mean number of road crashes in each district, taking into account both fixed and random effects. We specify the BYM2 model as follows:

$$y_i|\mu_i \sim \text{Poisson}(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 \times \text{Area}_i + \beta_2 \times \log(\text{Population})_i + \beta_3 \times \text{Station}_i + \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} = \tau_{\epsilon}^{-0.5}\left[(1-\phi)^{0.5}\boldsymbol{v} + \phi^{0.5}\boldsymbol{u}_*\right]$$

where y_i is the observed number of road crashes in district i; μ_i is the expected number of crashes in district i; β_1, β_2 and β_3 represent the slope coefficient for the covariates Area, log (Population) and Station respectively; u_i and v_i are the spatially structured and unstructured random effects, respectively; and ϵ_i is the residual error term. The spatially structured random effect, $\boldsymbol{u}_*$, accounts for the dependence between neighboring districts whereas the unstructured random effect, $\boldsymbol{v}$, accounts for unobserved heterogeneity and over-dispersion.

Five distinct neighborhood structures were chosen to serve as the spatial network representation. The initial two structures were derived from the queen's structures of first and second orders, each exhibiting a sparsity of 2.4% and 7.4%, respectively. The three remaining structures were determined by employing cut-off distances with threshold values of 75.2 km, $1.5 \times (75.2)$ km, and $2 \times (75.2)$ km, resulting in sparsity levels of 11.0%, 20.6%, and 31.6%, respectively. The minimum radius of 75.2 km was selected to ensure that the spatial field was fully connected for all neighborhood structures considered, as it represented the distance between the two farthest spatial units.

In fitting the BYM2 models, we utilized a PC prior with the following specifications for the hyper-parameters: $P\left(\tau_{\epsilon}^{-0.5} > 1\right) = 0.5$ and $P(\phi < 0.5) = 0.5$. For the fixed coefficients, we assigned normal priors with zero mean and a precision of 0.001.

Based on the results of the test of spatial autocorrelation using MI statistics (shown in Fig. 5), it was found that the first-order adjacency structure had the highest MI value, while the second order contiguity structure had the smallest p-value. Due to the sparsity of both structures, we rely on the MI statistics instead of the corresponding p-values. Based on the simulation inferences, we can infer that modeling with the first neighborhood structure is more likely to capture the appropriate level of spatial variability within the count outcomes compared to the other candidates. Hence, the estimates from the model that utilizes neighborhood structure 1 are preferable.

In Table 1, the fit statistics obtained after running all the BYM2 models with various neighborhood structure assignments are presented. Upon comparing the values of WAIC, DIC, MLIK, Sum of Squared Error (SSE), Mean Squared Prediction Error (MSPE), and CPO for each neighborhood structure representation, we notice similar values within each measure. Based on the fit statistics,

Table 1
Mean values of the hyper-parameters (standard deviations) and fit estimates for the BYM2 road crash model with varying neighborhood structures.

Graph	ϕ(s.d)	τ(s.d)	WAIC	DIC	MLIK	MSPE	CPO
1	0.2793 (0.1048)	0.5832 (0.0638)	2073.4	2130.9	−1534.6	0.833	1314.4
2	0.3517 (0.1269)	0.6012 (0.0623)	2073.5	2131.1	−1485.8	0.820	1315.8
3	0.2297 (0.0843)	0.5831 (0.0645)	2072.1	2130.4	−1561.3	0.824	1313.8
4	0.2758 (0.1038)	0.6089 (0.0619)	2071.2	2130.0	−1492.9	0.814	1314.5
5	0.3258 (0.1290)	0.6062 (0.0604)	2071.2	2130.1	−1479.6	0.804	1314.5

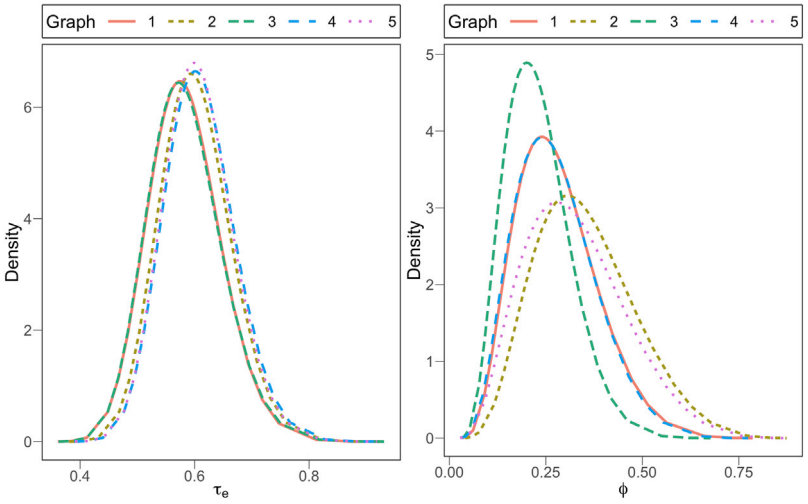


Fig. 6. Posterior distribution of the precision (left) and the mixing parameter (right) for the differently specified BYM2 models with varying neighborhood structures for road crash counts.

the densely populated spatial neighborhood 5 is a strong contender, as it yields the best-fit estimates. However, the CPO measure endorses the model with Structure 3, which is contradictory.

By examining the posterior distributions of the fixed effects for each of the prescribed neighborhood structures, it becomes apparent that the density of the two adjacency-based neighborhood structures are quite alike and different from the other three distance-based contiguity structures (refer to Fig. A.11). This implies that the choice of neighborhood structure specification may have an impact on fixed estimates, albeit not necessarily detrimental.

Similarly, Fig. 6 demonstrates the influence of the neighborhood structure effect on the estimation of the mixing parameter and precision. The density plots for τ_ϵ exhibit minimal variation due to the scaling of the spatial component, which ensures uniform effects of the prior density of the precision. Examining the posterior distribution of ϕ, we observe some differences, with peak densities for neighborhood structures 1 and 3 proposing smaller estimates of the mixing parameter compared to other structures. Additionally, as the mode approaches 0.5, the distribution becomes more symmetrical with a heavier right tail. Our preferred model produces the smallest value of ϕ, which represents approximately 28% of the random variability within the model.

Fig. 7 compares the structured effects $\boldsymbol{u}_*$ (top panel) and that of the combined random effects $\boldsymbol{\epsilon}$ (middle panel) as realized across the five neighborhood structures considered. From the plot, we observe that variations in the color pattern of both structured and combined random effects are largely influenced by the type of spatial network used in modeling. Thus, Structure 1 and Structure 2 which are both adjacency-based have similar patterns. Also, Structure 3, Structure 4 and Structure 5 which are all minimum-distance-based contiguity likewise share similar patterns with fewer variations. Meanwhile, the variations in color patterns between the adjacency-based structures and the minimum-distance-based structures are comparatively greater hence suggesting that significant differences may occur relative to our choice of neighboring structure.

A closer look at the plot of random effects for the preferred model (ie. the model under Structure 1) suggests that the positive effects as demonstrated by the map of structured random effects are linked to the coastal sectors and southern sectors in the country where there are more vehicular activities and increased human economic activities that warrant making use of the road network. Meanwhile, the negative effects are associated with the northern part of the country which is dominated by rural communities where industrialization is less and few human economic activities within the districts requiring fewer road users. We realize that the map of the combined random effect $\boldsymbol{\epsilon}$ is dominated by the unstructured component $\boldsymbol{v}$ since the fraction of the random effect attributed to the unstructured component is slightly above 70%.

Furthermore, the plot in Fig. 7 (lower panel) compares the predicted crash counts at the district level for all the models specified with different neighborhood structures. The brown regions represent an increasing number of incidents, while the ash color signifies a reduced number of crashes. Based on the color patterns, all the maps of predicted values look very similar, with the predicted

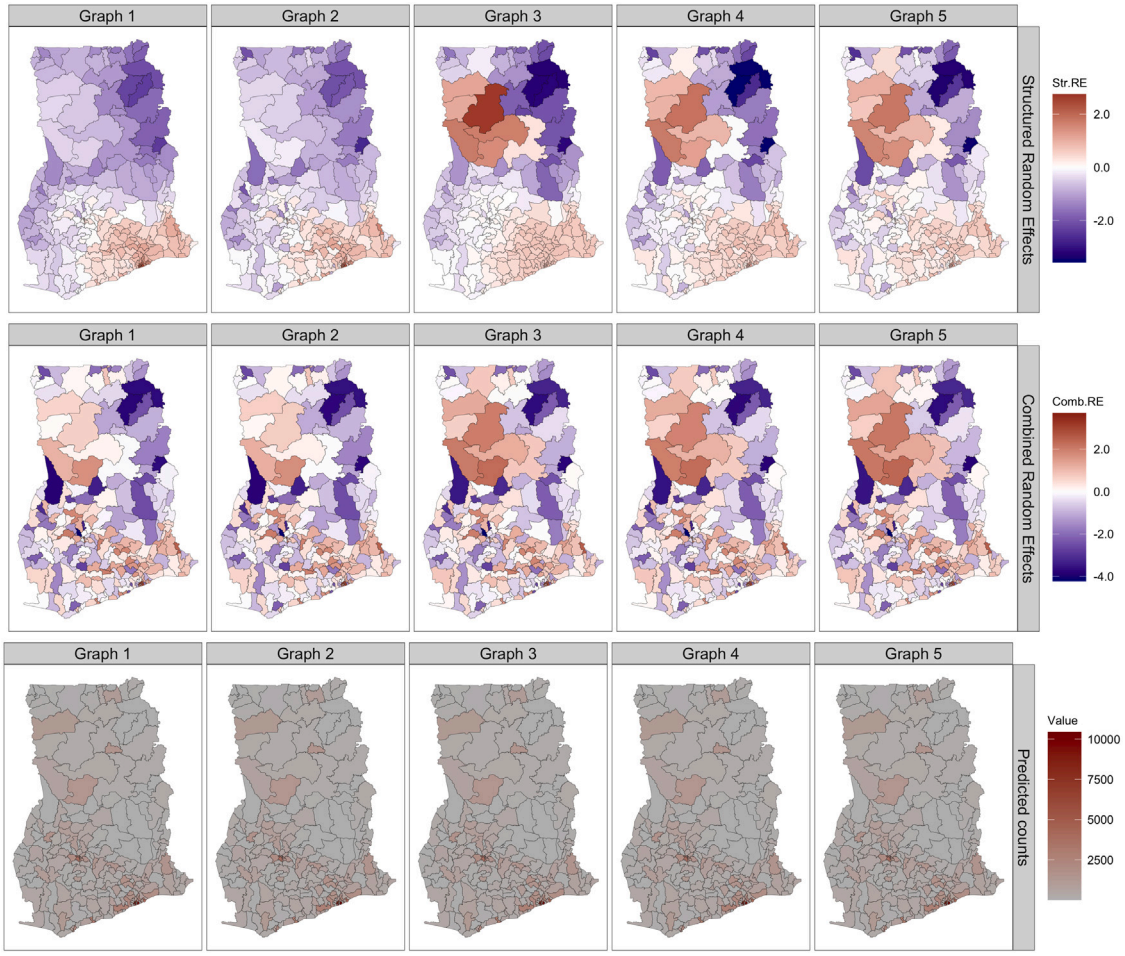


Fig. 7. Structured random effects $\boldsymbol{u}_*$ (top panel) and combined random effects $\boldsymbol{\epsilon}$ (middle panel) and predicted values (bottom panel) of crash counts for all BYM2 models specified with different neighborhood structures.

number of crashes being high for locations closer to the major cities in Ghana. This once again confirms the closeness observed in the predictive measures, as shown in Table 1. This suggests that while there may be significant variations within the random components depending on the type of spatial network structure utilized, these have minimal effects on prediction.

In Table 2, we present the fixed effect estimates for each of the covariates in the preferred model specification. Examining the fixed coefficients, we realize that the district's area in square kilometers appears to have a decreasing effect on the mean count, although its effect is not statistically significant at the 0.05 α-level. The coefficient for the log of Population meanwhile indicates a substantial increase in the expected number of crashes with population size. This suggests that every 1% increase in population size results in approximately a 1.54% increase in the expected crash count. The positive relationship emphasizes the role of population density in contributing to higher crash frequencies, likely due to increased vehicle and pedestrian traffic. The coefficient of 0.2336 for MTTD Stations also indicates that the presence of more MTTD stations is associated with an increase in crash frequencies. Specifically, for each additional MTTD station, the expected number of crashes increases by approximately 26.32%, a positive influence that is highly significant at the 0.05 α-level. This positive association probably reflects the fact that areas with more MTTD stations are also areas with higher traffic volumes and more enforcement activity, leading to a higher reported incidence of crashes.

Analysis of lip cancer incidents

We analyze data on lip cancer incidents among males in Scotland, sampled from 56 districts between 1975 and 1986. The data comprises aggregated lip cancer incidences linked to district centroids' coordinates where the incidents occurred during the study period. Lip cancer is potentially associated with sunlight exposure; thus, we use the percentage of individuals employed in the agricultural, fishery, and forestry sectors combined (PCAFF) as a covariate, representing individuals at risk of sun exposure due to their occupations. Additionally, the number of expected cases in each district is available as an offset term for modeling relative

Table 2
Estimates of coefficients from the BYM2 models for road crashes in Ghana for PC prior 1 and neighborhood structure 1.

Covariates	mean	sd	0.025 Q	0.975 Q
Intercept	−12.672	2.0016	−16.617	−8.7576
Area	−0.0001	0.0001	−0.0003	0.0000
Log.Pop.	1.5330	0.1738	1.1930	1.8754
Station	0.2336	0.0631	0.1099	0.3576

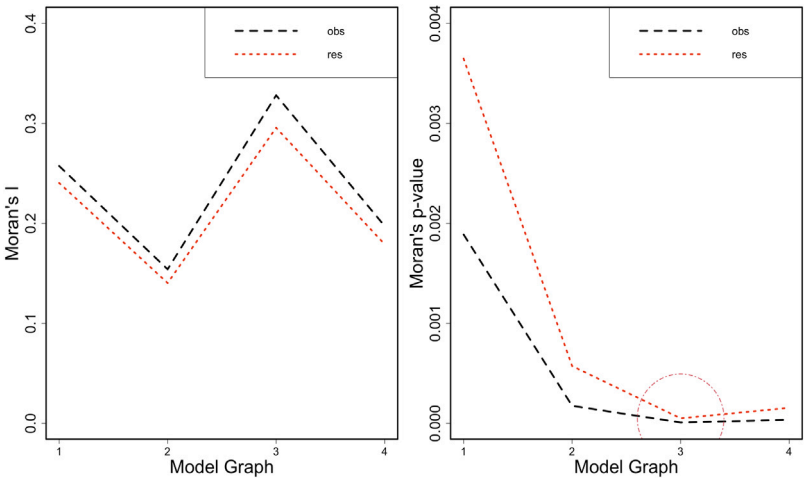


Fig. 8. Moran's I statistic (left) and p-value (right) for Lip cancer incidents and the residuals from the non-spatial model.

risk at specific locations. This dataset is stored in the **CARBayesdata** package in R and has been spatially modeled by various scholars [29–32].

From the dataset, the SIR for a given area i is simply the ratio of the observed value to the expected value in that region, i.e., $SIR_i = y_i/E_i$. The expected value E_i can be computed by indirect standardization, summing across all strata j within area i, the product of a standard measure of rate and the population size in the respective stratum $\left(\text{ie. } E_i = \sum_{j=1}^{m} r_j^{(s)} n_j^{(i)}\right)$.

We also attempt to model the SIRs using the BYM2 model with scaling and justify the choice of the neighborhood structure based on the findings from the simulations. Later, we demonstrate some of the discrepancies that may occur due to misspecification of the spatial structure. Here, we define the BYM2 model as follows:

$$y_i|\varphi_i \sim \text{Poisson}(E_i\varphi_i)$$

$$\log(\varphi_i) = \beta_0 + \beta_1 \times \text{PCAFF}_i + \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} = \tau_\epsilon^{-0.5}\left[(1-\phi)^{0.5}\boldsymbol{v} + \phi^{0.5}\boldsymbol{u}_*\right]$$

where φ_i stands for the relative risk in area i such that areas with high risks would be characterized by a high likelihood of $\varphi_i > 1$ and vice versa. β_0 also denotes the overall risk term, β_1 indicates the slope coefficient of PCAFF and $\log(E)$ denotes the expected count that incorporates different levels of exposure due to the size of the population assuming each area has its unique lip cancer rate. The terms τ_ϵ and ϕ represent the precision and mixing parameter of the random component $\boldsymbol{\epsilon}$ respectively. Specifically, τ_ϵ controls the marginal variance of the combined random effect and ϕ, the fraction of the random effect associated with spatial variation. The random component $\boldsymbol{\epsilon}$ which captures extra variability within the observed data is decomposed into a structured scaled component $\boldsymbol{u}_*$ and an unstructured random effect $\boldsymbol{v}$. The structured component is modeled as Intrinsic Gaussian Markov Random Fields (IGMRF) and the unstructured component is a Gaussian white noise.

We further defined four neighborhood structures to represent the spatial network underpinning the study. The first two structures were based on first and second-order contiguity with 8.2% and 25.1% nonzero entries respectively. The last two neighborhood structures were based on cut-off distances with structures 3 and 4 having radii of 50 km and 100 km with 14.6% and 38.6% nonzero elements respectively. The minimum radius of 50 km was arbitrary hence rendering some of the areal units isolated. We also assigned PC priors for the hyper-parameters, representing naive priors such that $P\left(\tau_\epsilon^{-0.5} > 1\right) = 0.5$ and $P(\phi < 0.5) = 0.5$. For the fixed effects, we assigned non-informative priors.

Table 3
Mean values of the hyper-parameters (standard deviations) and fit estimates for the BYM2 Lip cancer model with varying neighborhood structures.

Graph	ϕ(s.d)	τ(s.d)	WAIC	DIC	MLIK	MSPE	CPO
1	0.7759 (0.1911)	2.9588 (0.8946)	298.0	303.8	−141.1	0.156	180.2
2	0.4952 (0.2512)	3.0430 (0.8823)	304.8	310.1	−135.0	0.143	199.1
3	0.8931 (0.1127)	2.8391 (0.9041)	291.8	297.4	−142.3	0.137	185.6
4	0.7655 (0.2110)	3.0523 (0.8964)	299.8	305.6	−136.4	0.125	223.7

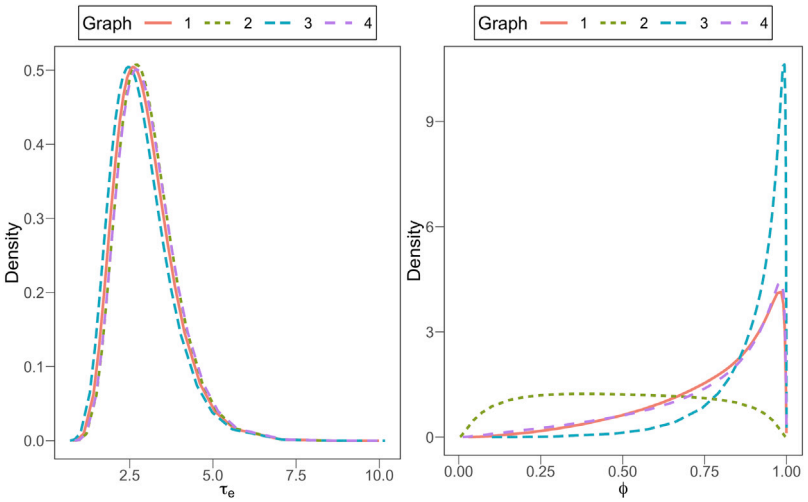


Fig. 9. Posterior distribution of the precision (left) and the mixing parameter (right) for the differently specified BYM2 models with varying neighborhood structures for Lip cancer incidents.

Table 4
Estimates of coefficients from the BYM2 models for Lip cancer incidents for PC prior 1 and neighborhood structure 1.

Covariates	mean	sd	0.025 Q	0.975 Q
Intercept	−0.3506	0.1311	−0.6122	−0.0964
PCAFF	0.0578	0.0125	0.0332	0.0825

A study of the Moran's test of spatial autocorrelation supports the use of the neighborhood structure 3 over the others since amongst the residual plots, Structure 3 has the largest MI statistic and at the same time, the smallest MI p-value thus indicating that structure 3 is most efficient at capturing the right proportions of residual spatial dependency as compared with the other neighborhood structures (see Fig. 8). Hence inferences on the spatial effects should be comparatively reliable under the model that utilizes structure 3 which is in support of our earlier decision.

Table 3 displays the estimates of WAIC, DIC, MLIK, MSPE and CPO for the relative risk of lip cancer under the various neighborhood structures specified. From the fit statistics, we observe the smallest WAIC and DIC values for the model with Structure 3. Judging from the CPO estimates, the smallest value occurred under the model specified with neighborhood structure 1. On the contrary, the MSPE estimates tend to favor the model with structure 4, which is much denser. Meanwhile, the MLIK is in support of the structure 2 thus contradicting the other model fit measures.

Focusing on the posterior distribution of the fixed effects (see Fig. A.12), we notice slight variations in the measure of center and spread with this variation slightly increasing between the contiguity-based structures and the distance-based structures. Thus this further suggests that the fixed effects are more resistant to variations in neighborhood structures. Therefore, misspecification of the spatial structure may slightly bias the fixed effects even though these errors may not be catastrophic.

Also, Fig. 9 illustrates the effects of the neighborhood structure on the posterior distribution of the hyper-parameters. The plot of τ_ϵ displays more resistance to the effect of neighborhood structure assignments as compared to that of ϕ. For ϕ, we observe greater variations in the posterior distributions with structure 3 appearing more stable and peaking the highest. This stability is reflected in Table 3 where the estimates for the precision term τ are close but the mixing parameter ϕ shows more variation with the highest estimate recorded for structure 3.

Fig. 10 maps over each district, the structured and combined random effects, and the posterior mean of lip cancer relative risk estimates across the utilized neighborhood structures. From the plot of structured effects, we see similarities between structures 1 and

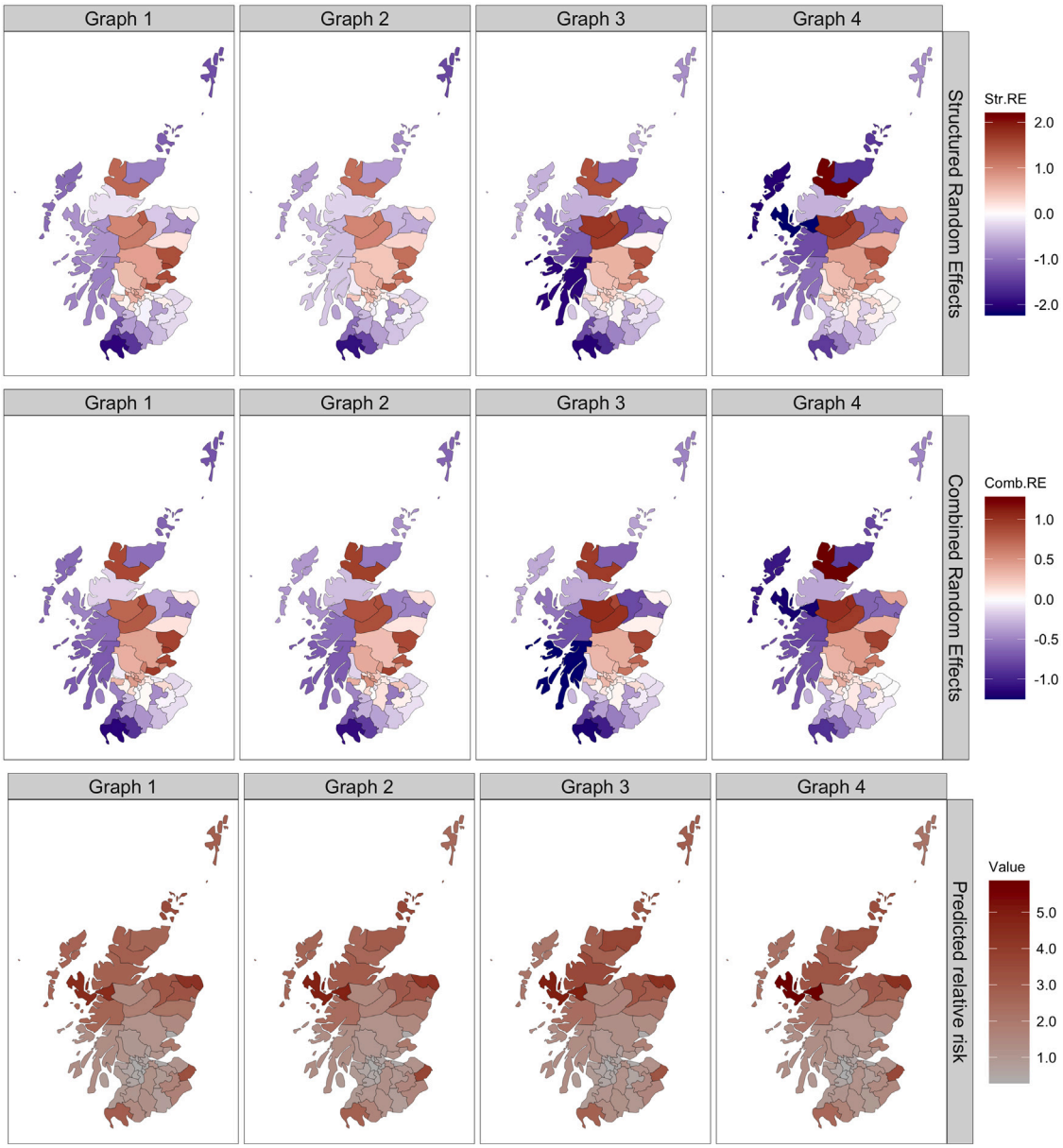


Fig. 10. Structured random effects $\boldsymbol{u}_*$ (top panel) and combined random effects $\boldsymbol{\epsilon}$ (middle panel) and predicted values (bottom panel) of Lip cancer incidents for all BYM2 models specified with different neighborhood structures.

2 which are from the same family of neighborhood structure. Likewise, we notice similarities in color patterns between structures 3 and 4 which are both distance-based. On the other hand, when contrasting models with different families of neighborhood structure, the differences are much more recognizable. Consequently, these graphical representations lead us to suggest that the variations in model outputs are primarily a result of the specification of the type of neighborhood structure. The plot of combined random effects seems dominated by the structured component $\boldsymbol{u}_*$ than that of the unstructured component $\boldsymbol{v}$. From the plot of predicted values, we see very close similarities in the pattern of lip cancer relative risk between the plots in the presence of neighborhood structure variations. We observe that districts in the northern part of Scotland are mostly associated with a high risk of lip cancer while those in the central belt of Scotland, UK are associated with a low risk of lip cancer.

From the preferred model, we observe an intercept of approximately -0.3474 with $(-0.6037, -0.0976)$ as the corresponding 95% credible interval. Likewise, the slope coefficient of PCAFF was about 0.0577 and the corresponding 95% credible interval was realized as $(0.0335, 0.0818)$. Hence this indicates that PCAFF increases the risk of lip cancer and that these two fixed components are

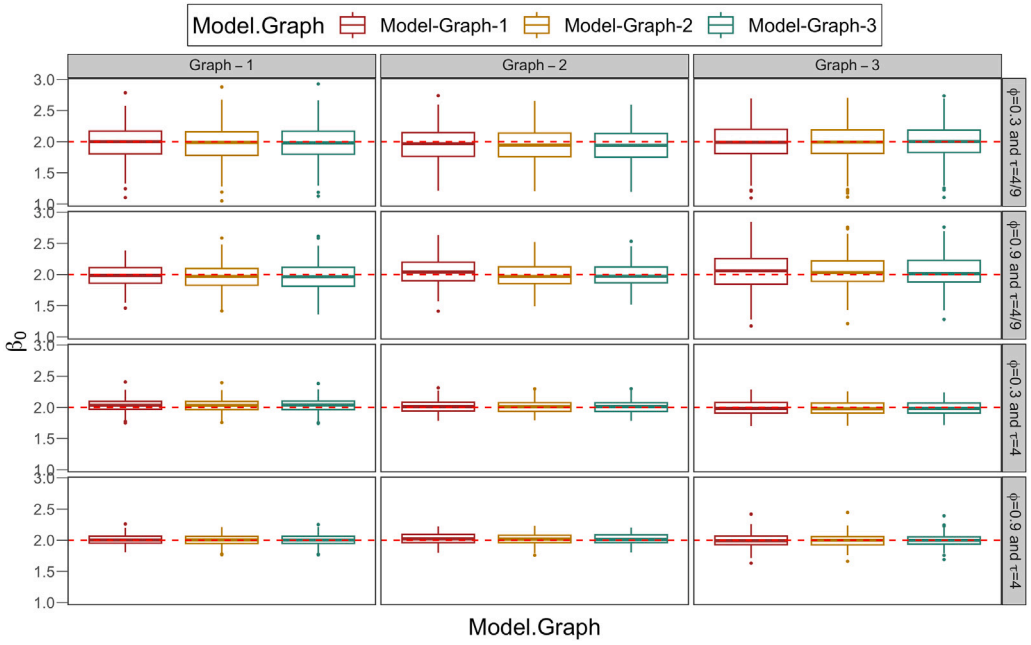


Fig. A.1. Distribution of posterior mean of β_0 from simulated results with misspecified neighborhood structures. The horizontal dashed line signifies the true parameter value. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

significant at the 0.05 α−level. Thus for every unit increase in the percent of PCAFF, we expect the relative risk of lip cancer to also increase by 5.95% assuming all other influential factors are held constant (see Table 4).

Discussion of results from the empirical studies

Among the neighborhood structures tested for the crash data, the first-order adjacency contiguity structure was the most informative for analyzing crash counts, as indicated by the highest MI value and the smallest p-value. Similarly, for the lip cancer data, the minimum distance with a radius of 50 km provided the best representation of the spatial network compared to the other neighborhood structures considered. Our analysis further revealed that the combined random effect in the crash model was primarily driven by the unstructured component, while in the lip cancer risk model, it was dominated by the spatial random component. The fixed effects in both empirical cases were relatively insensitive to neighborhood structure specification. Additionally, we observed greater variation in the values of the mixing parameter compared to the precision term when different spatial structure specifications were applied.

In interpreting the fixed effects of the lip cancer model, we observed that for every one-unit increase in the percentage of individuals working in the agricultural, fishery, and forestry sectors, there is a 5.95% increase in the relative risk of lip cancer incidents, assuming all other factors remain constant. In the crash model, population size has a strong and statistically significant positive effect on road crashes; a 1% increase in population leads to approximately a 1.53% increase in crashes. This outcome is expected, as larger populations typically result in increased road usage, higher vehicle density, and more interactions between drivers and pedestrians, all of which elevate the likelihood of road crashes. Similarly, an increase in the number of reporting stations is associated with a 23.36% rise in the mean crash count, indicating that areas with more crashes tend to have more police stations due to a higher need for traffic enforcement and public safety measures. Although the district's size appears to have a reducing effect on the expected crash count, its impact is negligible at the 0.05 alpha level.

Furthermore, we observed that variations in neighborhood structure definitions do not significantly affect predictions in either empirical study. We primarily attribute this to the values observed for both the spatial control term and the precision parameter. Based on the model structure, the impact of such misspecification on prediction is more likely to be significant in cases of very low precision and high spatial dependencies, particularly in real-life situations where high accuracy in prediction and forecasting is crucial. As a recommendation, the government and road agencies in Ghana should prioritize the improvement and enforcement of traffic regulations especially in urban areas and surrounding districts to reduce the number of crashes especially in such locations where economic activity is significant.

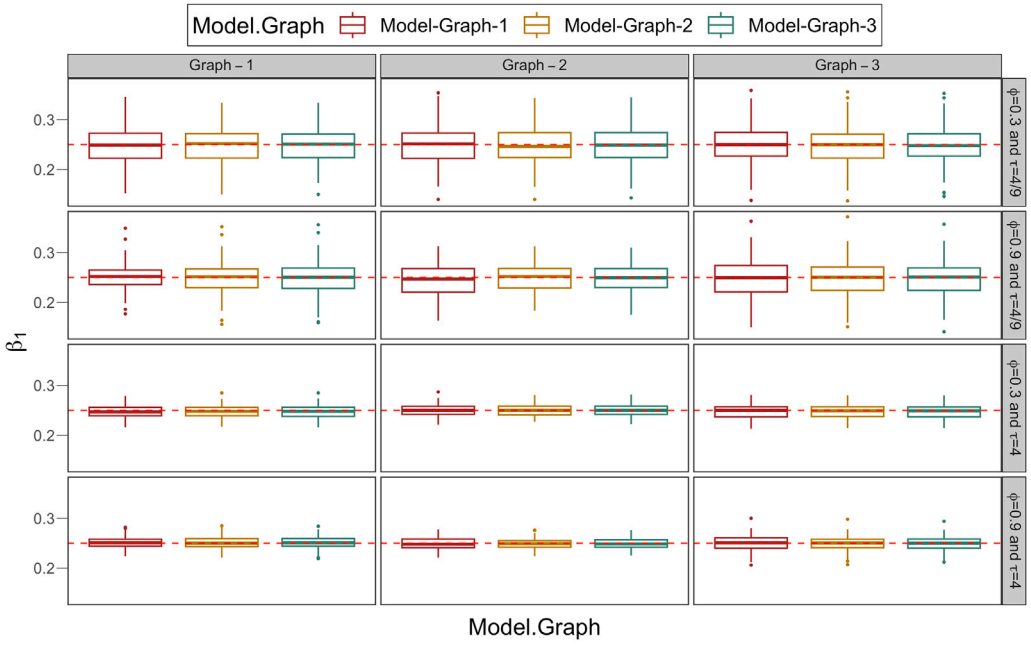


Fig. A.2. Distribution of posterior mean of β_1 from simulated results with misspecified neighborhood structures. The horizontal dashed line signifies the true parameter value. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

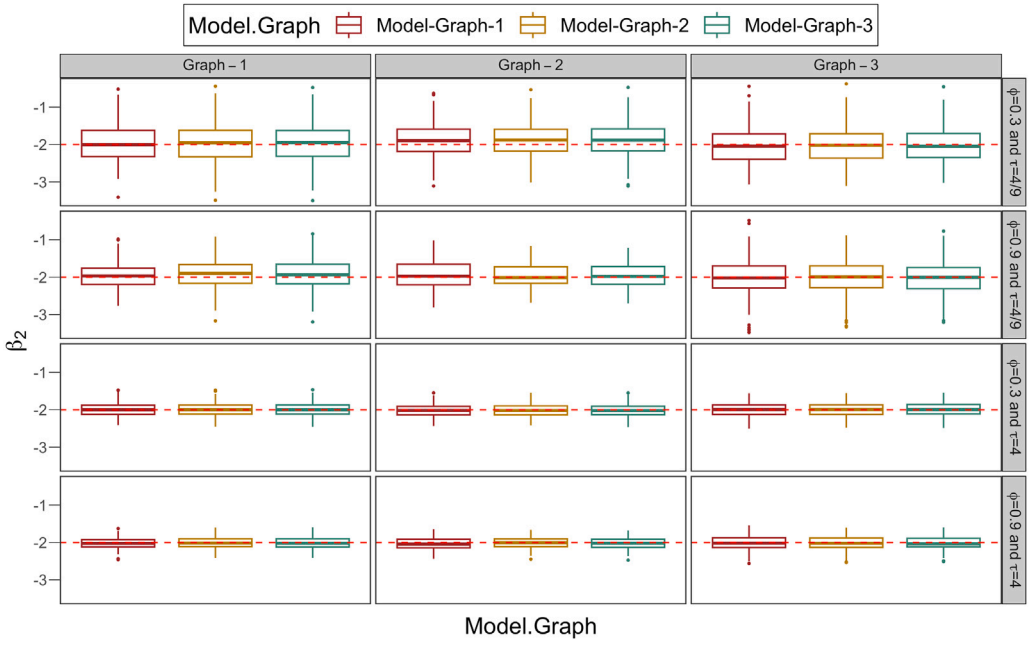


Fig. A.3. Distribution of posterior mean of β_2 from simulated results with misspecified neighborhood structures. The horizontal dashed line signifies the true parameter value. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

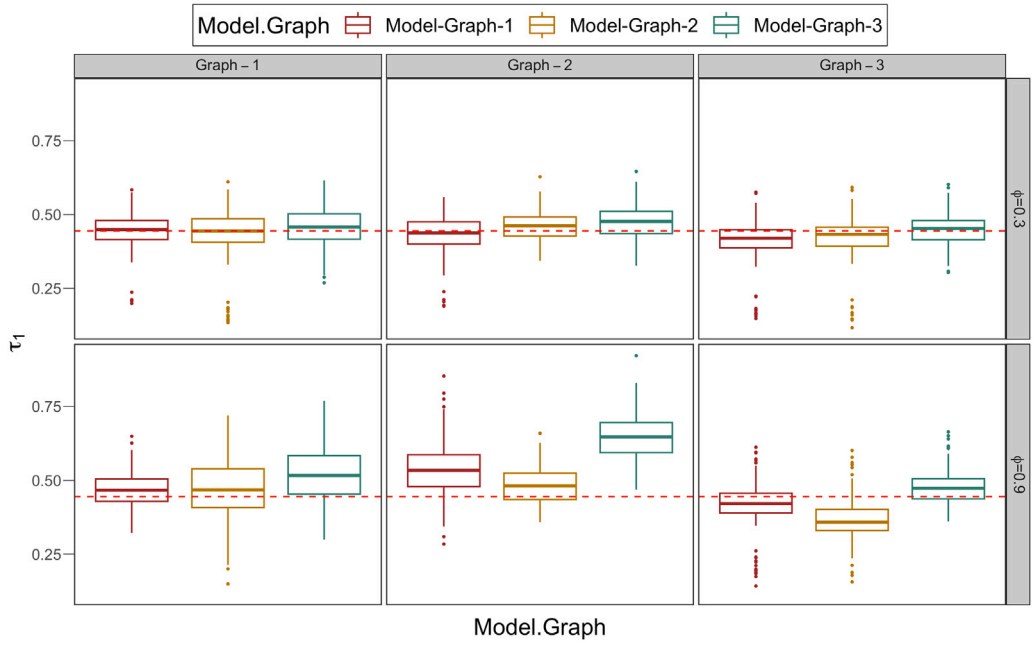


Fig. A.4. Distribution of posterior mean of τ_1 from simulated results with misspecified neighborhood structures. The horizontal dashed line signifies the true parameter value. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

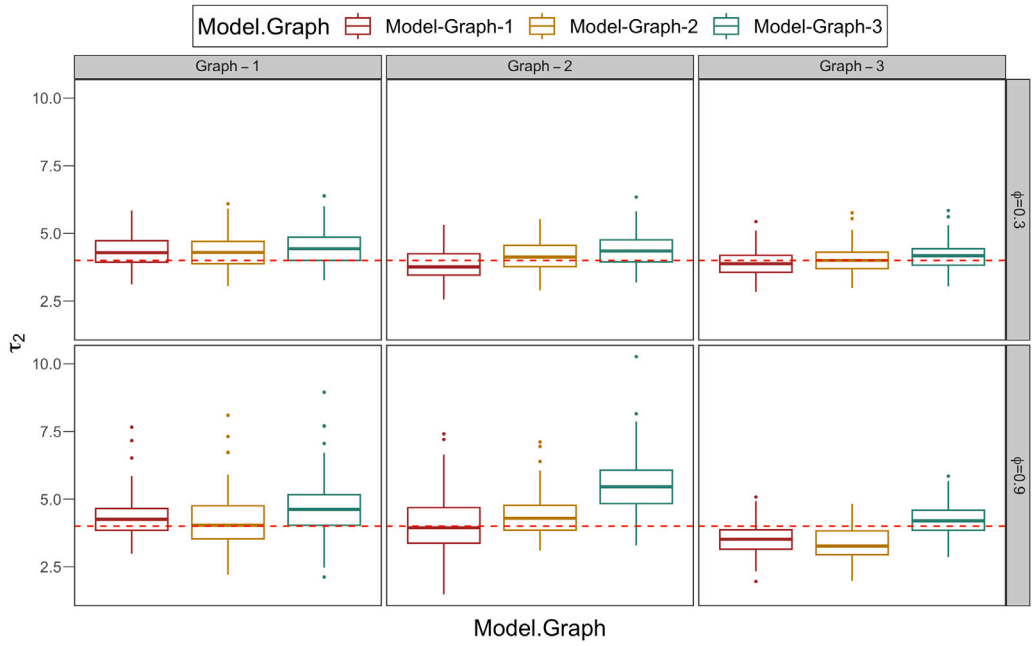


Fig. A.5. Distribution of posterior mean of τ_2 from simulated results with misspecified neighborhood structures. The horizontal dashed line signifies the true parameter value. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

Conclusion

In this study, we investigated the impact of misspecifying the spatial structure in BYM2 model framework through a simulation study and the application to two empirical data. The study by Duncan et al. (2017) is the closest to this study, although they examine the effects of different spatial weights by comparing several weighting matrices fitted to empirical data [2].

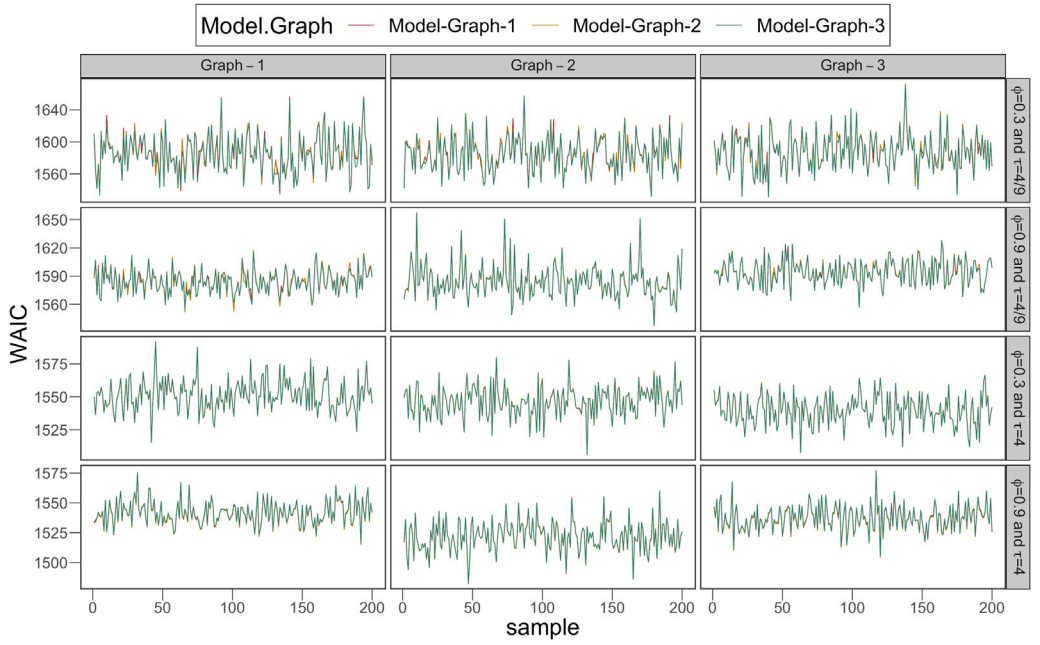


Fig. A.6. Plot of WAIC estimates from the models with neighborhood structure misspecification. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

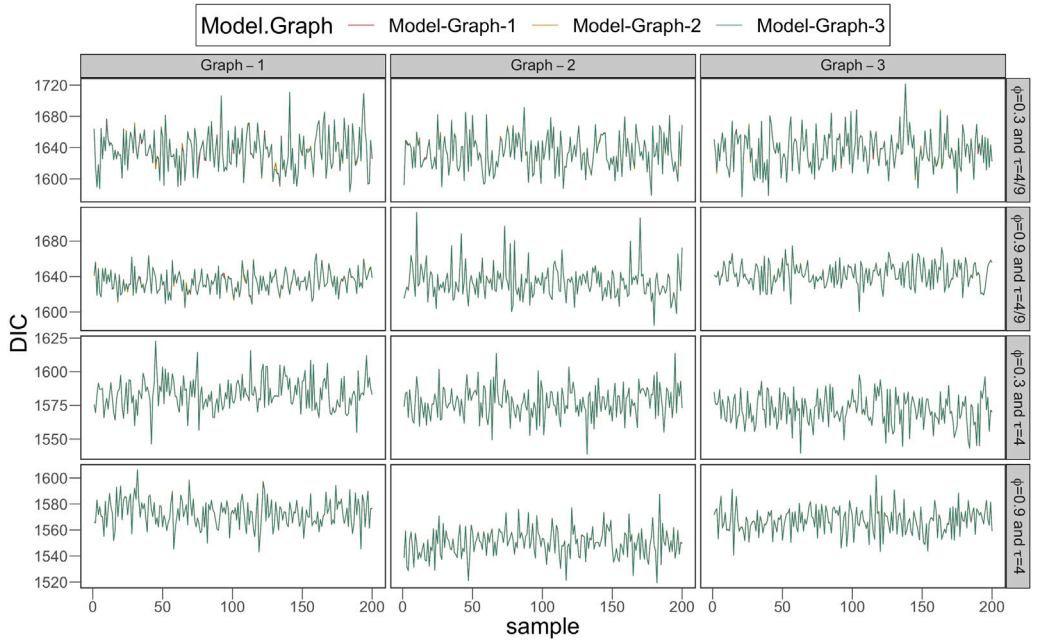


Fig. A.7. Plot of DIC estimates from simulated model with neighborhood structure misspecification. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

In our simulation study, we selected various neighborhood structures for districts of Ghana to meet our objectives, taking into account sparsity and computational efficiency. Our findings revealed that while neighboring structure misrepresentations significantly affected the mixing parameter and precision term, the fixed effects were less susceptible to such errors. Therefore, in comparison to the fixed effects, a wrongly specified neighborhood structure is more likely to introduce errors and distort the

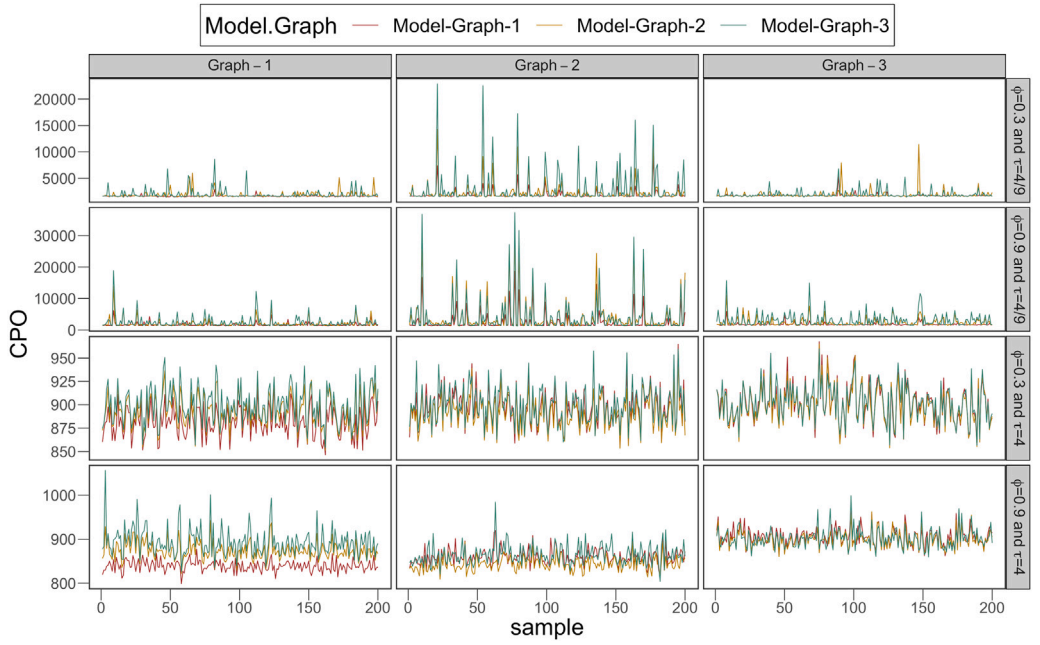


Fig. A.8. Plot of CPO estimates from simulated model with neighborhood structure misspecification. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

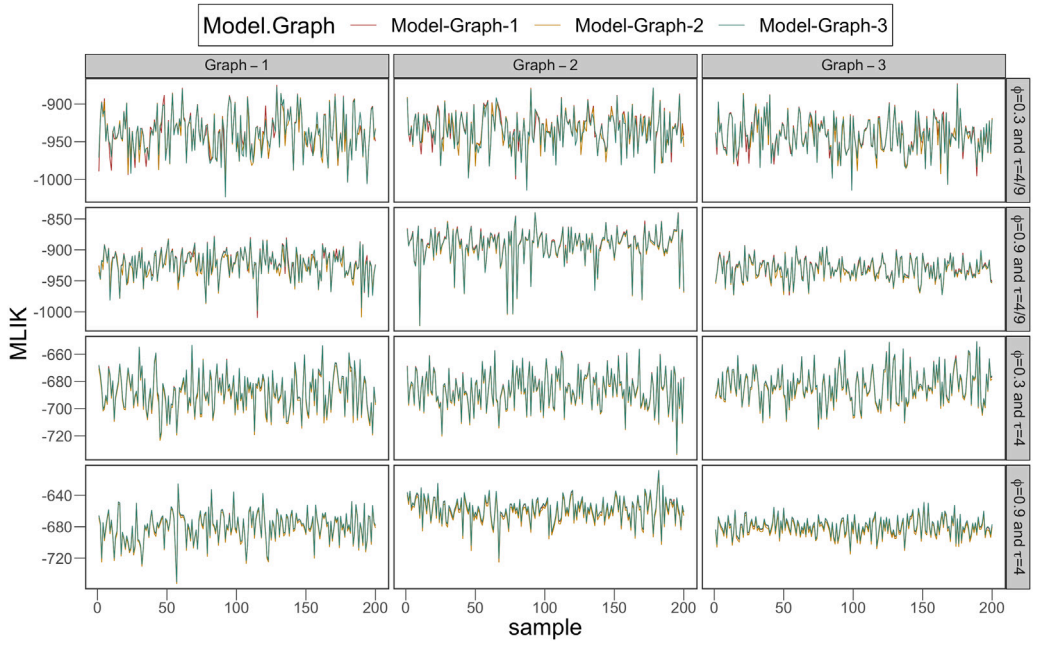


Fig. A.9. Plot of MLIK estimates from simulated model with neighborhood structure misspecification. Model-Graph-1 represents neighboring structure 1, Model-Graph-2 represents neighboring structure 2, and Model-Graph-3 represents neighboring structure 3.

estimates of the hyper-parameters. This emphasizes the importance of specifying accurate spatial structures for spatial models [3]. The study also indicated that under-specifying the true spatial network could lead to a loss of credible spatial information, resulting in an underestimated mixing parameter and precision term, hence an overly fitted spatial smooth for the random components. This problem is more severe when spatial autocorrelation is very low, regardless of the true value of the precision parameter. In contrast, if the spatial structure is over-specified, additional noise is introduced, resulting in an overestimation of the mixing and

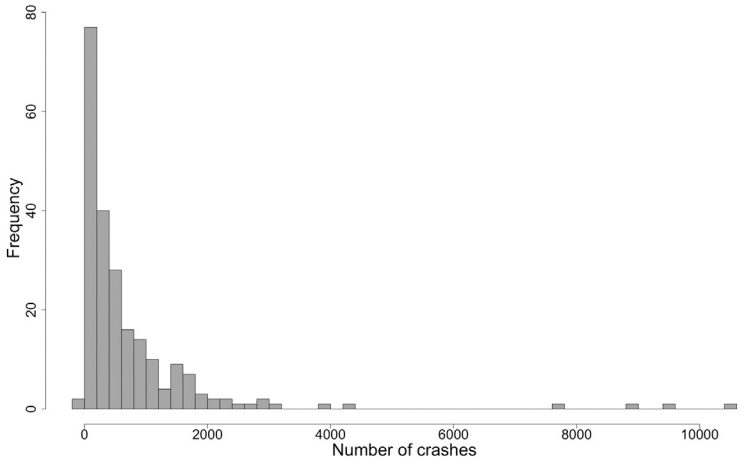


Fig. A.10. Histogram of the number of road crashes in Ghana accumulated at the district level.

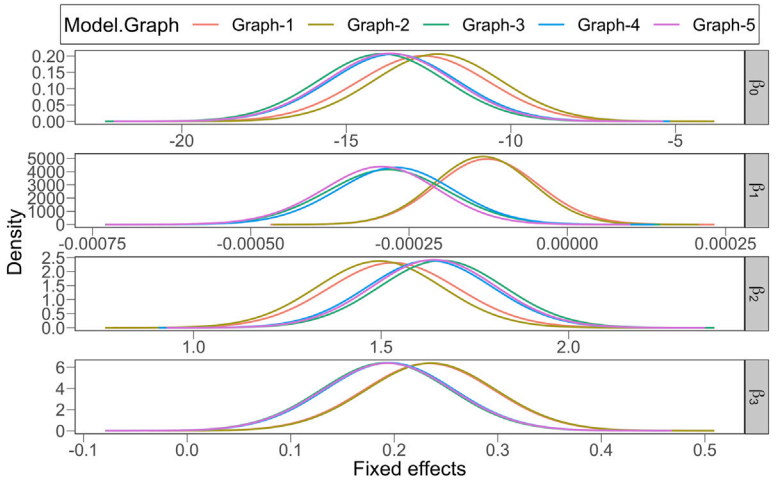


Fig. A.11. Posterior distribution of fixed effects for the differently specified BYM2 models with varying neighborhood structures for road crash counts.

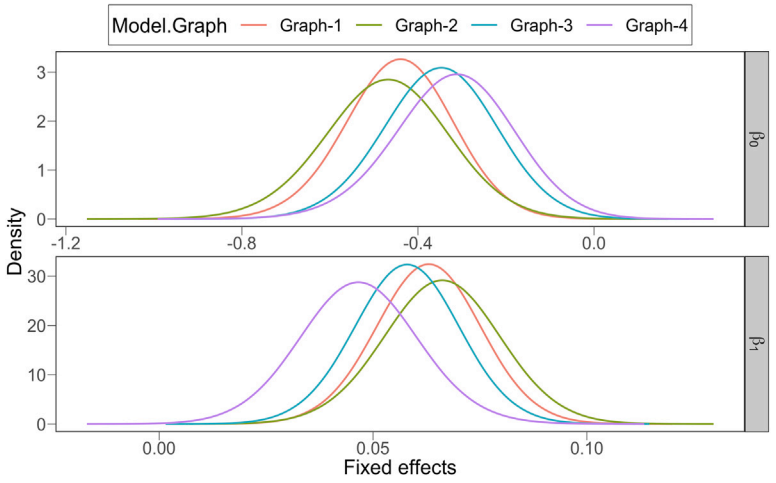


Fig. A.12. Posterior distribution of fixed effects for the differently specified BYM2 models with varying neighborhood structures for Lip cancer incidents.

precision parameters. In this case, the spatial component is under-fitted, showing less random variability than expected and creating a false impression of a good overall fit. Under-specifying or over-specifying the spatial weights can lead to the mistaken treatment of structured and unstructured random effects, resulting in identifiability issues [33]. However, in real-life applications, the true spatial network is often more complex than it appears, making strictly under-specified or over-specified cases less common.

Though measures such as WAIC, DIC, CPO, and MLIK can be helpful for model assessment and selection, particularly in cases where the models have varying formulations and complexities, they may not have the ability to identify the spatial model with the best weight structure from multiple options. Therefore, if the goal of the study is to select a model with the most suitable neighborhood structure that minimizes biases and errors, relying solely on the aforementioned measures would not be advisable. This observation is collaborated by [34] who covered the limitations and contexts in which DIC may or may not be useful for model comparison, particularly for complex hierarchical models, while [28] discussed WAIC and other criteria that may fall short in models with spatial dependencies or varying neighborhood structures. Instead, the MI test of spatial autocorrelation can play a vital role in identifying the true spatial network and thus can be appropriately used in identifying the best neighborhood structure among all other structures that represent the true spatial network under study. Our study demonstrated that the sparser the spatial weight matrix, the greater the likelihood that the MI statistic would accurately identify the optimal neighborhood structure. Conversely, when identifying the optimal neighborhood structure among less sparse options, it is advisable to rely on the MI p-values. The most suitable spatial weight matrix will consistently exhibit the smallest p-value among all other neighborhood structures under consideration.

Further investigations were conducted where the BYM2 model was fitted to road crash counts in Ghana over 255 areal units utilizing five different spatial neighborhood structures, and lip cancer incidence rates in Scotland, UK, across 56 areal units. The crash data exhibited relatively low spatial dependency while the lip cancer data demonstrated a high level of spatial dependency hence mirroring the scenarios studied in our simulation studies. The posterior distributions obtained were compared, and the results of the empirical cases validated the inferences drawn from the simulation exercise.

This study has a limitation in that it does not consider the impact of misspecifying the spatial structure as the number of area units and replications increase, which is a computationally intensive task. With the advancements in computational power, researchers may need to consider exploring this direction in future studies [35]. Additionally, experiments could be conducted to examine the impact of biases introduced by the misspecification of neighborhood structure in the face of extreme random effects. To further explore this topic, researchers could explore merging the misspecification of spatial weights with spatially correlated covariates with or without measurement errors to study the behavior of the spatial model under such conditions. Additionally, they could employ a different spatial CAR or specify a spatio-temporal random field [36] and analyze how an erroneously specified space–time covariance structure affects regression outputs over increasing samples.

CRediT authorship contribution statement

Ernest Somua-Wiafe: Conceptualization, Data curation, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Richard Minkah:** Supervision, Writing – review & editing, Validation. **Kwabena Doku-Amponsah:** Methodology, Supervision, Writing – review & editing. **Louis Asiedu:** Methodology, Writing – review & editing. **Edward Acheampong:** Validation, Writing – review & editing. **Samuel Iddi:** Conceptualization, Supervision, Methodology, Validation, Writing – review & editing.

Ethical consideration

This study used secondary and simulated data where no personal, sensitive or identifying information was extracted.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We sincerely acknowledge the immense contribution from Mr. Joshua Larley, a Chief Technical Officer at Building and Road Research Institute (BRRI), who assisted with accessing and securing the road accident data and proved useful in compiling and cleaning the crash data.

Supplementary appendix

See Figs. A.1–A.12.

References

[1] A. Earnest, G. Morgan, K. Mengersen, L. Ryan, R. Summerhayes, J. Beard, Evaluating the effect of neighbourhood weight matrices on smoothing properties of conditional autoregressive (CAR) models, Int. J. Health Geogr. 6 (1) (2007) 54.

[2] W.E. Duncan, M.N. White, K. Mengersen, Spatial smoothing in Bayesian models: a comparison of weights matrix specifications and their impact on inference, Int. J. Health Geogr. 16 (1) (2017) 1–16.

[3] R. Assunção, E. Krainski, Neighborhood dependence in Bayesian spatial models, Biom. J. 51 (5) (2009) 851–869.

[4] Á. Briz-Redón, A. Iftimi, J.F. Correcher, J. De Andrés, M. Lozano, C. Romero-García, A comparison of multiple neighborhood matrix specifications for spatio-temporal model fitting: a case study on COVID-19 data, Stoch. Environ. Res. Risk Assess. 36 (1) (2022) 271–282.

[5] T. Betz, S.J. Cook, F.M. Hollenbach, Bias from network misspecification under spatial dependence, Polit. Anal. 29 (2) (2021) 260–266.

[6] S. Juhl, The sensitivity of spatial regression models to network misspecification, Polit. Anal. 28 (1) (2020) 1–19.

[7] R. Arambepola, T.C.D. Lucas, A.K. Nandi, P.W. Gething, E. Cameron, A simulation study of disaggregation regression for spatial disease mapping, Stat. Med. 41 (1) (2022) 1–16.

[8] R.J.G.M. Florax, S. Rey, The impacts of misspecified spatial interaction in linear regression models, in: New Directions in Spatial Econometrics, Springer, Berlin, 1995, pp. 111–135.

[9] T.E. Smith, Estimation bias in spatial models with strongly connected weight matrices, Geogr. Anal. 41 (3) (2009) 307–332.

[10] T. Rüttenauer, Spatial regression models: A systematic comparison of different model specifications using Monte Carlo experiments, Sociol. Methods Res. 51 (2) (2019) 1–32.

[11] X. Wang, J. Yang, C. Lee, Z. Ji, S. You, Macro-level safety analysis of pedestrian crashes in Shanghai, China, Accid. Anal. Prev. 96 (2016) 12–21.

[12] D. Simpson, H. Rue, A. Riebler, G.T. Martins, H.S. Sø rbye, Penalising model component complexity: A principled, practical approach to constructing priors, Statist. Sci. 32 (1) (2017) 1–28.

[13] J. Chen, J.J. Song, J.D. Stamey, A Bayesian hierarchical spatial model to correct for misreporting in count data: Application to state-level COVID-19 data in the United States, Int. J. Environ. Res. Public Health 19 (6) (2022) 3327.

[14] P.A. Gao, J. Wakefield, Smoothed model-assisted small area estimation of proportions, Can. J. Stat. 52 (2) (2024) 337–358.

[15] M. Victoire, P.F. Laís, M.S. Alexandra, G.C. Oswaldo, A Bayesian hierarchical model for disease mapping that accounts for scaling and heavy-tailed latent effects, 2024.

[16] N. Asmarian, S.M.T. Ayatollahi, Z. Sharafi, N. Zare, Bayesian spatial joint model for disease mapping of zero-inflated data with R-INLA: A simulation study and an application to male breast cancer in Iran, Int. J. Environ. Res. Public Health 16 (22) (2019) 4460.

[17] J. Besag, J. York, A. Mollie, Bayesian image restoration, with two applications in spatial statistics, Ann. Inst. Statist. Math. 43 (1) (1991) 1–20.

[18] A.S. Fotheringham, C. Brunsdon, M. Charlton, Geographically Weighted Regression: The Analysis of Spatially Varying Relationships, Wiley, UK, 2002.

[19] A. Getis, J.K. Ord, The analysis of spatial association by use of distance statistics, Geogr. Anal. 24 (3) (2010) 189–206.

[20] J.M. Ver Hoef, E.P. Erin, B.H. Mevin, M.H. Ephraim, F. Marie-Josee, Spatial autoregressive models for statistical inference from ecological data, Ecol. Monograph. 88 (1) (2018) 36–59.

[21] L. Anselin, G. Morrison, A. Li, K. Acosta, Hands-on spatial data science with R, 2020.

[22] A. Riebler, S.H. Sørbye, D. Simpson, H. Rue, An intuitive Bayesian spatial model for disease mapping that accounts for scaling, Stat. Methods Med. Res. 25 (4) (2016) 1145–1165.

[23] H. Rue, S. Martino, N. Chopin, Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2) (2009) 319–392.

[24] P.A.P. Moran, Notes on continuous stochastic phenomena, Biometrika 37 (1) (1950) 17–23.

[25] A. Hubin, G. Storvik, Estimating the marginal likelihood with integrated nested Laplace approximation (INLA), 2016, http://dx.doi.org/10.48550/arXiv.1611.01450, arXiv E-Prints.

[26] V. Gómez-Rubio, H. Rue, Markov chain Monte Carlo with the integrated nested Laplace approximation, Stat. Comput. 28 (5) (2018) 1033–1051.

[27] S. Watanabe, Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, J. Mach. Learn. Res. 11 (2010) 3571–3594.

[28] A. Gelman, J. Hwang, A. Vehtari, Understanding predictive information criteria for Bayesian models, Stat. Comput. 24 (6) (2014) 997–1016.

[29] D. Clayton, J. Kaldor, Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, Biometrics 43 (3) (1987) 671–681.

[30] J. Wakefield, Disease mapping and spatial regression with count data, Biostatistics 8 (2) (2007) 158–183.

[31] M. Morris, K. Wheeler-Martin, D. Simpson, S.J. Mooney, A. Gelman, C. DiMaggio, Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan, Spat. Spatiotemporal Epidemiol. 16 (47) (2019) 1–16.

[32] D. Lee, M. Durbán, Smooth-car mixed models for spatial count data, Comput. Statist. Data Anal. 53 (8) (2009) 2968–2979.

[33] R. Haining, Spatial Data Analysis: Theory and Practice, Cambridge University Press, 2003.

[34] D. Spiegelhalter, N. Best, B. Carlin, A. Van Der Linde, Bayesian measures of model complexity and fit, J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (2002) 583–639.

[35] A.O. Finley, S. Banerjee, B.P. Carlin, spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models, J. Stat. Softw. 19 (4) (2007) 1–24.

[36] N. Cressie, C.K. Wikle, Statistics for Spatio-Temporal Data, John Wiley & Sons, 2011.