UNMEASURED SPATIAL RISK IN DISEASE MODELING A TOPOLOGICAL DATA ANALYSIS APPROACH

HASHAN FERNANDO

Abstract. The opioid epidemic is a complex public health crisis that has caused immense damage and devastation in the United States. The complexity of this epidemic is driven by various socioeconomic and demographic factors, such as unemployment, disability, poverty, overprescription, and accessibility of opioids. In spatial epidemiology, researchers incorporate these factors as covariates within Bayesian spatial models to estimate the areal-level overdose risk. While standard Bayesian spatial models successfully stabilize subregional-level (small spatial units) risk estimates by addressing issues such as data sparsity, they struggle to estimate the risk at the regional level (large spatial units). At the regional level, standard models tend to over-smooth the risk estimates, thereby masking sharp transitions in risk among neighboring areas. To address this challenge, we use topological data analysis (TDA), a collection of tools that use concepts from algebraic topology to identify robust topological features such as connected components and loops. We extract these topological features from subregional-level covariate data and integrate the resulting summaries into regional-level Bayesian spatial models. By providing region-specific topological summaries, our approach reduces the uncertainty in risk estimates, thereby allowing the model to effectively identify sharp transitions in risk among neighboring areas. Using real-world opioid mortality and social vulnerability data from West Virginia and Tennessee, we empirically evaluated the performance of our proposed framework. The results show that the TDA-integrated approach captures the topological summaries necessary to identify sharp transitions in risk among neighbors. As a result, our method not only performs as well as current models but also reduces the uncertainty in regional estimates and accurately identifies sharp transitions in risk.

1. Introduction

The opioid epidemic has resulted in an unprecedented number of overdose deaths in the United States [20]. This epidemic has progressed in a series of phases based on the availability of the drugs, beginning with prescription opioids, followed by heroin, and gradually escalated to catastrophic levels by fentanyl and other synthetic opioids [12]. Instead of spreading uniformly, the use of each drug followed a unique geographic path, starting in specific, densely populated areas before progressively spreading to impact communities throughout the United States [25]. This pattern of drug-specific spread indicates that the opioid epidemic propagates like a contagion, influenced by a complex interaction of socioeconomic factors that affects different communities in different ways [12, 16]. Therefore, understanding where overdose risk concentrates, how it evolves, and which regions remain most vulnerable demands a spatially informed analytical framework. Disease mapping provides this exact framework to study the spread of the epidemic, thereby allowing us to identify the root causes of opioid use and opioid overdose.

One of the most effective tools for disease mapping is Bayesian spatial modeling [6]. Bayesian spatial models are specifically designed to address the challenges of "small area estimation," which

Key words and phrases. Topological Data Analysis, Opioid Epidemic, Bayesian Spatial Modeling.

involves estimating disease risk in small geographic areas where data are often sparse [28]. To mitigate the instability of sparse data, Bayesian spatial models use spatial smoothing, where they adjust unstable risk estimates toward local neighborhood averages [38, 6]. In particular, this spatial risk adjustment is mathematically modeled using a Conditional Auto-Regressive (CAR) prior, a mathematical framework that explicitly accounts for spatial dependence among adjacent geographic units [8]. Despite the demonstrated effectiveness of CAR priors in producing stable small-area risk estimates, their usefulness ultimately depends on the availability of small-area data, a condition that cannot always be met in practice. For example, outcome data such as overdose deaths are often reported at the broader regional level (e.g., county level), which has large spatial units, due to privacy regulations that require masking small counts to protect anonymity [37]. In contrast, covariate data may be gathered at the subregional level (e.g., census tract level), which has small spatial units [20]. To resolve this spatial-level mismatch, researchers aggregate subregional-level covariate data to match regional-level outcome data, allowing them to estimate disease risks at the broader, regional level [10, 6]. However, estimating disease risks at the regional level using Bayesian spatial models introduces a fundamental issue. In particular, the CAR component in standard Bayesian spatial models imposes a uniform level of smoothness across the study region while ignoring the sharp transitions among adjacent neighbors [27]. This leads to an over-smoothing of risk estimates, reducing their certainty and reliability. Given this limitation, there remains a need for a framework that can improve regional-level Bayesian spatial modeling to reduce uncertainty in their risk estimates and identify sharp transitions in risk among neighbors.

To address the limitation of regional-level Bayesian spatial modeling, we propose an integrated framework that combines topological data analysis (TDA) [18], a collection of tools that use concepts from algebraic topology to extract robust structural properties, with hierarchical Bayesian modeling. In our framework, we identify topological features, such as connected components, using subregional-level covariate data and integrate the resulting topological summaries into regional-level Bayesian spatial models. By including topological summaries of subregional-level covariate data directly into the disease mapping process, our framework achieves two key goals: (1) it provides region-specific topological information to the models; (2) it decreases the uncertainty in regional-level risk estimates, ultimately allowing the models to identify sharp transitions in risk among neighbors. We demonstrate the potential of this approach by modeling opioid overdose risk, transforming underutilized subregional-level covariate data into a helpful asset in disease mapping to achieve precise risk estimates.

1.1. **Related Work.** The Besag, York, and Mollié (BYM) model is widely used in public health for disease mapping because it provides reliable disease risk estimates by accounting for measurable risk factors and spatial dependence in a disease [3, 40]. This model incorporates spatial dependence as a "spatial random effect" within hierarchical Bayesian modeling. In addition, it recognizes that not all spatial dependence is the same, as some are shared among neighbors while others are specific to a single location. To reflect this distinction, the BYM model divides the spatial effect into two components: a structured effect and an unstructured effect. The unstructured effect accounts for localized random noise, whereas the structured effect accounts for spatial dependence among adjacent regions [9]. The structured effect is often modeled using a CAR prior distribution. This CAR prior represents the structured effect for a region by averaging the effects of adjacent regions [8, 36]. On the other hand, the unstructured effect is modeled using a Gaussian prior distribution, representing localized random noise [39]. Moreover, the relative contributions of these two effects are controlled by separate precision parameters. The unstructured precision parameter controls the variance of the unstructured effect, while the structured precision parameter controls the variance

of the structured effect. Although the BYM model introduces a novel approach to represent spatial dependence, it has certain limitations.

In particular, the BYM model has two well-known limitations. First, the BYM model has an identifiability issue: it cannot distinguish the percentages of spatial dependence allocated to the structured effect or to the unstructured effect [28, 15]. Second, and more importantly, the BYM model exhibits a scaling issue: properties of the graph structure, such as the number of nodes and shape, rather than the precision parameter, control the variance of the structured component. Therefore, the overall variance of the spatial random effect, which is a combination of the variances of both structured and unstructured effects, depends on the graph structure of the study region. As a result, the graph structure implicitly increases or decreases the overall variance of the spatial random effect, leading to inconsistent risk estimates across different spatial applications with different graph structures [39, 41].

In response to these limitations, multiple extensions of the BYM framework have been formulated [28, 15]. To overcome the identifiability issue, alternative formulations were developed by Leroux et al. [28] and Dean et al. [15]. While both approaches introduce a mixing parameter to control the percentage of spatial dependence, both approaches differ in mechanism: Leroux et al. proposed a single random effect with a precision matrix accounting for properties of both structured and unstructured spatial effects [28]; Dean et al. preserved the two random effects and mathematically reparameterized the spatial random effect to control their overall variance using a single precision parameter [15]. These earlier models, however, failed to address the scaling issue. To resolve this issue, building on the reparameterization proposed by Dean et al. [15], Riebler et al. [39] introduced the BYM2 model. This formulation introduces a mathematically scaled structured component to make sure that the precision parameter accurately reflects the overall variance of the spatial random effect, regardless of the graph structure of the study region. Despite these advancements in identifiability and scaling, the BYM framework still struggles to estimate risk using large spatial units at the regional level. At the regional level, sharp risk transitions can exist among neighboring areas, yet the CAR component imposes a uniform level of smoothness across the entire study region. This smoothing ignores the sharp risk transitions among neighbors, ultimately leading to uncertain and potentially misleading risk estimates [27].

The key to identifying these sharp risk transitions lies in equipping the model with region-specific information that distinguishes genuine spatial patterns from random noise. TDA addresses these challenges by offering a collection of tools that use concepts from algebraic topology to extract robust structural properties [17]. Among these tools, persistent homology stands out as a primary method for distinguishing true structural properties from noise [17, 43, 35]. Recent studies have shown that persistent homology can be used to detect spatial patterns in data. For example, Byers et al. showed that persistent homology can be used to identify non-random clustering in geospatial data [11], and Limbeck and Rieck extended this to find spatial patterns in high-dimensional transcriptomics data [30]. Additionally, Feng et al. used persistent homology to identify spatial voting patterns of the 2016 US presidential election that were overlooked by standard methods [19], while Hickok et al. successfully tracked spatial anomalies in COVID data using persistent homology [21]. However, despite the proven ability of persistent homology to capture complex spatial patterns, it has not yet been employed in the disease risk mapping process. Motivated by its effective application in spatial graphs, we aim to further develop persistent homology as a robust framework to extract region-specific robust spatial patterns from subregional-level covariate data, ultimately decreasing the uncertainty in disease mapping models to identify sharp transitions in disease risk.

1.2. **Contribution.** The primary contribution of this work is a novel TDA-Bayesian framework that extracts topological summaries from subregional-level covariate data and integrates them into regional-level Bayesian spatial modeling, thereby mitigating uncertainty in regional-level risk estimates and identifying sharp transitions in disease risk. We empirically validate this methodology by modeling opioid mortality in West Virginia and Tennessee, employing Markov Chain Monte Carlo (MCMC) sampling via the PyMC library [1] to demonstrate that our topology-integrated model yields better risk estimates compared to standard Bayesian approaches.

The outline of the paper is as follows. In Section 2, we present the background, beginning with disease risk mapping and the fundamentals of TDA. In Section 3, we present the comprehensive methodology of our proposed approach. Section 4 presents the results of our synthetic data studies, real-world case studies, and sensitivity analysis. Section 5 presents the discussion and the conclusion.

2. Background

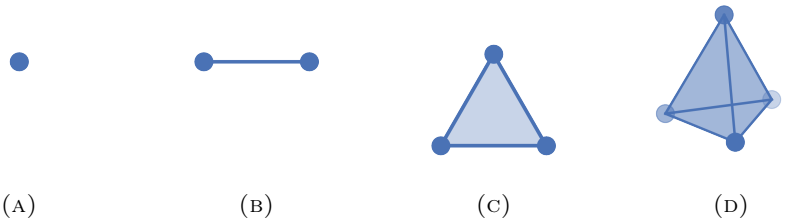
2.1. **Data Overview.** The opioid overdose mortality data used in this study are obtained from the CDC WONDER database for 2018 [20]. In this dataset, multiple causes of death are classified according to the International Classification of Diseases, 10th Revision (ICD-10), which provides the framework for identifying overdose cases. We focus on codes X40–X44 (unintentional), X60–X64 (suicide), X85 (homicide), and Y10–Y14 (undetermined intent), which together capture opioid-related mortality. Mortality rates are calculated per 100,000 population, based on data recorded in death certificates filed in all 50 states and the District of Columbia.

Socioeconomic data for US counties and census tracts are obtained from the CDC/ATSDR SVI [20] for 2018. We utilize 15 social vulnerability index (SVI) variables, including unemployment, poverty, and disability, to capture community-level conditions. These variables are designed to identify communities that experience hazards and other stressors, and prior studies have shown that these socioeconomic factors have a significant impact on opioid overdose deaths [4, 16, 5, 7]. The county and census tract boundary files used for generating spatial adjacency are sourced from the US Census Bureau.

2.2. **Disease Risk Mapping.** In standard spatial epidemiology, rare and non-infectious diseases are typically modeled using a Poisson distribution [31]. To apply this modeling strategy to our study region A, we first divide the study region A into n mutually exclusive spatial units, $A_1, \ldots, A_n$. Within each spatial unit i, we track the observed count of deaths y_i and calculate the expected count of deaths E_i. This E_i is calculated by specifically accounting for the population size of area i. A general form of the log-linear model is given by

$$\begin{aligned} y_i|\theta_i &\sim \text{Poisson}(E_i\theta_i), \quad i = 1, \ldots, n, \\ \eta_i &= \log(\theta_i), \end{aligned}$$

where θ_i denotes the relative risk of area i, and η_i is its log-transformation. Estimating these risks independently using standard methods presents two fundamental issues. First, it disregards the impact of measurable socioeconomic and environmental risk factors, such as poverty and disability, on disease outcomes. Second, and more importantly, independent risk estimations result in highly unstable risk estimates in sparsely populated areas, where even small changes in the observed death counts can lead to significant change in the risk estimates, resulting unreliable risk estimates with wide confidence intervals. To address both challenges simultaneously, spatial Bayesian modeling framework offers a natural and well-established solution for disease mapping [31, 9, 26]. Spatial Bayesian models generate far more robust results by shifting extreme risk estimates toward the



(A) (B) (C) (D)

FIGURE 1. Visual representation of simplices. The figure displays (A) a 0-simplex, (B) a 1-simplex, (C) a 2-simplex, and (D) a 3-simplex.

local neighborhood average while accounting for the risk factors [8, 28, 15]. In standard spatial Bayesian modeling, the log-relative risk η_i is typically represented as

$$\eta_i = \xi + x_i^T \beta + b_i.$$

In this formulation, ξ sets the overall risk level and β captures the effect of m fixed-effect covariates x_i^T, while b_i captures the spatial random effect. Incorporating the random effect b_i is a critical step to account for the impact of unobserved factors [39]. To capture these factors, our baseline approach relies on a modern reparameterization of the standard BYM [9] framework known as the BYM2 model [39]. The reparameterized spatial random effect of the BYM2 model is given by

$$b_i = \frac{1}{\sqrt{\tau_b}}(\sqrt{1-\phi}v_i + \sqrt{\phi}u_i^*),$$

where τ_b is a precision parameter controlling the total variance of b_i, the scaled structured spatial component u_i^* captures dependence among adjacent neighbors using an intrinsic CAR prior, while a Gaussian error term v_i accounts for any remaining unstructured, localized noise. The mixing parameter $\phi \in [0, 1]$ controls the relative contribution of each component. Although the BYM2 model offers a principled approach to spatial risk estimation, a critical limitation becomes apparent when applying it to large spatial units at the regional level. At this spatial level, risk levels can change sharply between neighboring areas. However, the BYM2 model tends to over-smooth these risk transitions among neighbors, masking sharp boundaries in risk [27]. To overcome this limitation, we incorporate topological summaries of the subregional-level covariate data into regional-level models. By providing additional topological information to the regional-level models, we reduce the uncertainty in risk estimates, thereby allowing the models to detect sharp transitions in risk among neighbors. Motivated by this objective, section 3.4 introduces a novel reparameterization of the spatial random effect.

2.3. **Simplicial Complex and Homology.** Formally, a finite k-simplex is a structure formed by connecting $k+1$ independent points $\{p_0, p_1, \ldots, p_k\}$, and any simplex formed by a subset of these points is called a face. A simplicial complex is a set consisting of nodes (0-simplices), edges (1-simplices), triangles (2-simplices), tetrahedra (3-simplices), and their higher-dimensional polytopes (see fig. 1) [18]. For a finite collection of simplices to form a valid simplicial complex, they must satisfy the gluing conditions:

(1) Inclusion conditions: Every face of a simplex must also be included in the simplicial complex.

(2) Intersection conditions: The intersection of any two simplices in the simplicial complex is either empty or is a face of both.

Once a valid simplicial complex has been generated, we investigate its structure using homology, a tool in algebraic topology [18, 34, 43]. Homology counts topological features such as the number of connected components (0-dimensional topological features), loops (1-dimensional topological features), and voids (2-dimensional topological features) [43, 34].

3. Methods

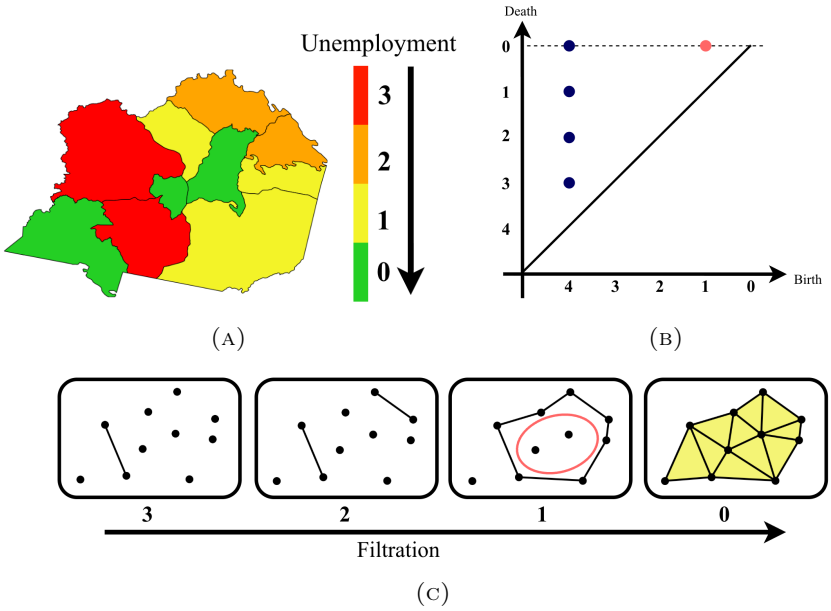
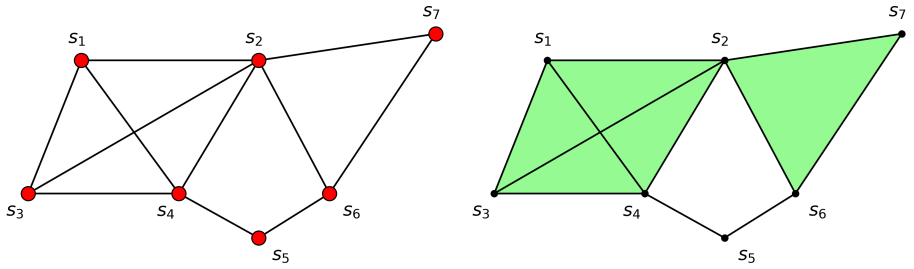


Figure 2. Overview of the proposed framework. Panel (A) demonstrates the spatial distribution of the unemployment variable in subregional spatial units, (B) demonstrates computing persistent homology by tracking the evolution of topological features, and (C) demonstrates defining filtration based on the unemployment values and geographical adjacency.

3.1. **Modeling Subregional Connectivity as a Spatial Graph.** To model subregional-level spatial data as a spatial graph, consider a spatial region A_i partitioned into N distinct subregions, creating a set $S = \{s_1, \ldots, s_N\}$. Each unit $s_r \in S$ represents a geometric region $R(s_r)$ (a polygon or multipolygon) and features an associated covariate value $x(s_r)$. Following Feng and Porter [19], we model these units as nodes of a spatial graph $G = (S, \mathcal{E})$. In this graph, edges are defined by shared geographical boundaries rather than spatial proximity, a construction known to create more accurate topological representations [19, 33]. To be more specific, we connect two nodes s_e and s_j only if their geometric boundaries $R(s_e)$ and $R(s_j)$ share any boundary points, including corners. The edge set is formally defined as

$$\mathcal{E} = \{(s_e, s_j) \in S \times S \mid \ s_e \neq s_j \text{ and } s_e \text{ adjacent to } s_j\}.$$



(A) (B)

FIGURE 3. (A) Example of a graph G; (B) Graph G's complex of cliques $= < s_1, s_2, s_3, s_4 > \cup < s_2, s_6, s_7 > \cup < s_4, s_5 > \cup < s_6, s_5 >$.

Building upon this spatial graph foundation, we next describe the process of extracting topological features from the graph G.

3.2. **Persistent Homology from Spatial Data.** To transition from the spatial graph G to a framework capable of capturing topological features, we adapt the topological approach recently applied to spatial transcriptomics by Limbeck and Rieck [30]. We begin by transforming the spatial graph G into a clique complex K by "filling in" every complete subgraph (clique) in G (see fig. 3) [18]. After establishing the clique complex K, we use super-level set filtration with the covariate values directly as the filtration parameter δ to examine the spatial distribution of subregional-level covariate data $x(s_r)$. Beginning at the maximum covariate value across all subregions, we gradually decrease δ, at each threshold, defining the active vertex subset as $S_\delta = \{s_r \in S \mid x(s_r) \geq \delta\}$. The corresponding subcomplex K_δ is then constructed by including all simplices $\zeta \in K$ whose vertices are entirely contained in S_δ. In this super-level set filtration, reducing the filtration value adds new simplices without removing any that already exist because of the monotonicity of this construction. As a result, we obtain a finite sequence of nested subcomplexes represented as

$$\emptyset \subseteq K_{\delta_1 = \max(x(s_r))} \subseteq \cdots \subseteq K_{\delta_L = \min(x(s_r))} = K,$$

where each subcomplex K_{δ_l} is associated with a filtration value δ_l for $l = 1 \ldots L$, forming a filtered simplicial complex [35]. To demonstrate this method in practice, fig. 2C shows a sequence of subcomplexes generated when applying this super-level set filtration to subregional-level unemployment data.

As this filtration progresses, instead of examining the homology of an individual K_{δ_l}, we examine the homology of the nested sequence. As a result, we monitor the emergence and disappearance of topological features throughout the nested sequence. We denote the emergence and disappearance of each topological feature using a pair of birth (w) and death (z) values. The filtration value δ_l at which a topological feature first appears (e.g., when a new connected component forms) is represented by the birth, and the filtration value δ_l at which the topological feature disappears is represented by the death. However, the particular mechanism of this disappearance varies according to the type of feature being examined. For instance, when two connected components merge at the filtration step l, the recently added component is considered to have died (known as the "elder rule"). Typically, any components that never die are assigned a death time of infinity ($z = \infty$) [17, 18].

The difference between death and birth represents the persistence of a topological feature. In the literature, researchers who used persistent homology ignored short-lived topological features as random noise, focusing primarily on long-lived features to understand underlying topological structures. However, this is not a universal rule; recent applications demonstrate that the distribution of these short-lived features actually contains vital information [2].

The birth and death filtration values of all topological features in the above filtration are represented in a persistence diagram (see fig. 2B), formally defined as a finite multi-set (a collection of points where each point can appear more than once) of points in the upper half of the first quadrant of $\mathbb{R}^2$ [17, 14].

3.3. **Topological Summaries.** However, since a persistence diagram contains a multi-set of birth-death pairs, it is challenging to examine from a statistical perspective. Standard statistical models typically require fixed-size vector representations, whereas persistence diagrams allow the number of features to vary for each observation. This mismatch necessitates transforming persistence diagrams into a more statistically beneficial representation. Therefore, we have chosen to include topological summaries in our study based on the findings of Byers et al. [11]. In particular, we consider the following statistics:

(1) Average Lifespan (AL) $= \frac{1}{T}\sum_{t=1}^{T}(z_t - w_t)$
(2) Average MidLife (AML) $= \frac{1}{T}\sum_{t=1}^{T}(z_t + w_t)/2$

Here, w_t denotes the birth, z_t denotes the death of a topological feature t, and T denotes the total number of topological features. Although the above statistics provide useful scalar summaries, aggregation is necessary to construct them. As a result of this aggregation, a degree of the original information is lost, potentially complicating the interpretation of the results. In fact, distinct diagrams might map to the same scalar, making it difficult to trace their topological meaning.

Despite this limitation, topological summaries effectively capture critical insights [30, 11]. For each covariate in the disease risk model, we therefore compute persistence diagrams and their topological summaries, focusing on connected components. We consider connected components because persistence pairs in the persistence diagrams detect peaks in spatial data [24]. During this computation, we utilize the minimum filtration value as a standard way to bound the death value of each persistence pair.

3.4. **Integrating Topological Summaries into Spatial Random Effect.** Small area disease risk estimation models, such as the BYM2 model, smooth their risk estimates based on the risk similarity of their neighbors. This smoothing stabilizes estimates in small areas, thereby overcoming issues such as data sparsity. However, these models struggle to estimate the risk at the regional level, with large spatial units, where disease risk estimates change sharply even among neighbors. Therefore, excessive reliance on smoothing results in unreliable risk estimates.

To address this, we propose the BYM-topological (BYM-T) model, which extends the BYM2 framework by incorporating a topological component. This component acts as a region-specific signal that integrates topological summaries of subregional-level covariate data. By offering region-specific topological summaries, we are reducing the uncertainty in risk estimates, thereby enabling the model to identify significant, sharp risk differences among neighboring areas. The log relative risk of the proposed model is given by

$$\eta_i = \xi + x_i^T\beta + b_i^{new},$$

where ξ represents the overall risk level, $x_i^T = (x_{i1}, \ldots, x_{im})$ represents a vector of m covariates with related coefficients $\beta = (\beta_1, \ldots, \beta_m)$, and b_i^{new} is the reparameterized spatial random effect.

This spatial random effect is given by

$$b_i^{new} = \frac{1}{\sqrt{\tau_b}}(\sqrt{\alpha_1}v_i + \sqrt{\alpha_2}u_i^* + \sqrt{\alpha_3}h_i),$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$. Here, u_i^* represents a scaled structured component, ensuring variance contributed by the graph structure is standardized to one [39], v_i represents the unstructured component, and h_i represents the aggregated topological summaries (AL) of covariates, defined as

$$h_i = \sum_{j=1}^{m} h_{ij},$$

where the topological component h_i is standardized such that its variance is equal to one.

The BYM-T model is a direct generalization of the BYM2 model. To see this, note that when $\alpha_3 = 0$, the BYM-T spatial random effect reduces to a two-component combination. This combination is identical in structure to BYM2, with α_1 serving as $1-\phi$ (unstructured weight) and α_2 serving as ϕ (structured weight), subject to the constraint $\alpha_1 + \alpha_2 = 1$. The introduction of the topological component h_i with weight α_3 provides a third pathway rather than forcing all unexplained spatial effects into either localized noise or neighborhood smoothing. All three components in the BYM-T model are scaled to unit variance, so α_1, α_2 and α_3 are directly interpretable as the proportion of total spatial random effect assigned to each source (see appendix B).

Intuitively, the unstructured component captures region-specific noise that has no spatial pattern, and the structured component captures neighboring spatial dependency. However, at the regional level, the topological component fills a gap that neither of these can address. The topological component captures region-specific spatial patterns in the subregional-level covariate data. For example, two counties can have identical county-level poverty rates (identical fixed covariate values) and identical neighbor connections yet have very different internal (census tract-level) spatial distributions: one with poverty concentrated in a single hotspot and another with poverty uniformly distributed. The topological summary h_i differentiates between these two counties, allowing the model to assign different risk estimates even when all other inputs are equal. The core novelty of the BYM-T model lies in its ability to transform typically discarded patterns in subregional covariate data into useful information, thereby reducing the uncertainty of risk estimates.

4. Numerical Experiments

4.1. **Synthetic Data.** To generate synthetic data, we first construct a geospatial lattice, that divides the entire geographic region into individual subregional spatial units. Then, for each subregional-level spatial unit, we generate synthetic covariate data. Specifically, we simulate spatially correlated data using the Spatial Autoregressive (SAR) process, defined as

$$x = \rho W x + \epsilon,$$

where ρ is the spatial autocorrelation parameter, W is the row-standardized adjacency matrix that represents the neighborhood structure, and ϵ is the normally distributed random error [29]. The resulting x values represent simulated subregional-level covariate data that can be interpreted as synthetic versions of the subregional-level SVI data. We generate synthetic covariate datasets with varying spatial autocorrelation while keeping other factors, such as the geographical structure and the regional covariate value (or aggregated subregional-level covariate data), constant. As shown in fig. 4A , these synthetic data appear as grid cells where color indicates the magnitude of the covariate values.

4.1.1. *Experimental Setup & Results.* Our objective in this synthetic data study is to compare regions that share the same regional covariate value but exhibit distinct spatial distributions in their subregional-level covariate data using topological summaries. To do this, we generate a set of datasets for five counties with the same geographical structure and the same regional covariate value, varying only the autocorrelation parameter ρ from –0.9 to 0.9 (fig. 4A). The autocorrelation serves as a statistical indicator of the spatial distribution of subregional-level covariate data, indicating the degree to which neighboring subregions group together (forming peaks) or disperse. Then we employ persistent homology to extract topological summaries from each dataset. Topological summaries are plotted against autocorrelation to assess their sensitivity to changes in the spatial distributions of subregional-level covariate data.

As illustrated in fig. 4B topological summaries, AML and AL serve as effective indicators of peaks of subregional-level covariate data. Although these topological summaries remain low for negatively autocorrelated data, they increase considerably as positive autocorrelation increases. This sensitivity is directly associated with the spatial distribution of the subregional-level covariate data. Specifically, at high positive autocorrelation ($\rho = 0.9$), the subregional-level covariate data forms several distinct peaks or hotspots, whereas at negative autocorrelation, the data remains dispersed (see fig. 4A).

It should be noted that, in real-world geographic settings, the number of neighboring regions can vary significantly. In contrast, our simulated maps maintain relatively consistent neighboring patterns, which is a limitation of our synthetic data study.

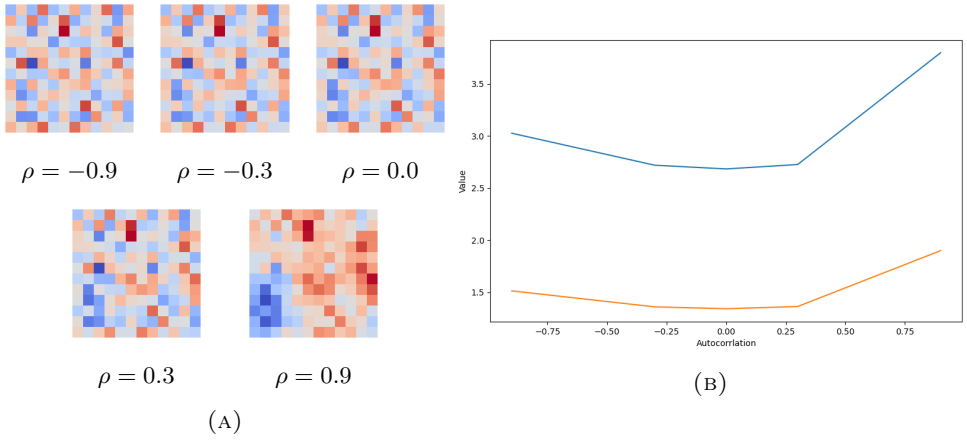


FIGURE 4. Left panel: Synthetic datasets showing varying levels of spatial autocorrelation from $\rho = -0.9$ to $\rho = 0.9$. Right panel: Corresponding topological summaries extracted from the synthetic datasets. The AML (blue line) and AL (orange line) illustrate the response to controlled changes in spatial autocorrelation.

4.2. **Real-World Studies.** Our study centers on West Virginia and Tennessee, two states significantly impacted by the opioid epidemic. West Virginia is completely included in the Appalachian Region, which is primarily rural. This environment makes it difficult to access healthcare and distribute resources. Tennessee faces similar difficulties, mainly in its Appalachian areas in the east.

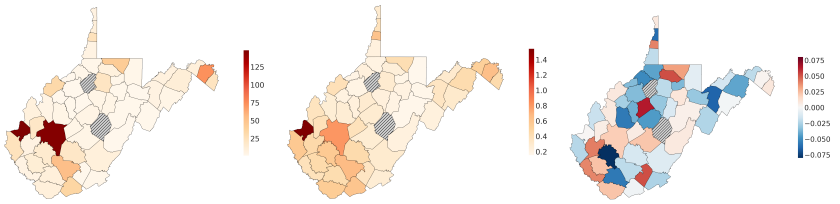
Both states consistently report higher opioid mortality rates than national averages, making them critical environments for evaluating new disease risk models [32, 22].

Administratively, West Virginia consists of 55 counties, and Tennessee of 95. However, data availability is constrained by CDC WONDER privacy protocols, which suppress mortality counts below ten. Due to these suppressed counts, two counties were excluded from the West Virginia models and four from the Tennessee models, resulting in final sample sizes of $n = 53$ and $n = 91$, respectively.

4.2.1. *Experimental Setup.* To evaluate the effectiveness of our proposed topological approach, we estimate the opioid overdose risk using both the proposed BYM-T model and the standard BYM2 model. In both frameworks, we use opioid overdose mortality counts as the response variable and SVI variables as independent fixed covariates.

Our outcome data and covariate data are collected at different spatial levels. Opioid mortality data are recorded at the county level, while SVI data are measured at the census tract level. For fixed covariates in both the BYM2 and BYM-T models, we aggregated census-tract SVI data to the county level to ensure a consistent unit of analysis. Additionally, in the BYM-T model, we utilize census-tract-level covariate data to calculate the topological summaries of the covariates, providing additional region-specific information to the model.

In the model specification in both the BYM2 and BYM-T models, we assign the following prior distributions to the hyperparameters: $\xi \sim \mathcal{N}(0, 1)$, $\beta \sim \mathcal{N}(0, 1)$, $\phi \sim \text{Beta}(1, 1)$, $1/\sqrt{\tau_b} \sim \text{HalfNormal}(0.39)$, $(\alpha_1, \alpha_2, \alpha_3) \sim \text{Dirichlet}(1, 1, 1)$. We compare the BYM-T and BYM2 models using the expected log predictive density from leave-one-out cross-validation (ELPD_{LOO}) [42]. We implemented our models in Python using the PyMC probabilistic programming framework [1]. The PyMC framework supports flexible model specification and efficient Bayesian inference using modern samplers. For our inference engine, we use nutpie, an engine designed to work seamlessly with PyMC's NUTS (No-U-Turn Sampler) to improve computational performance. We assess convergence by running multiple independent chains and monitoring the $\hat{R}$ statistic, the effective sample size (ESS), trace plots, and rank plots (see appendix A). Throughout all experiments, the computational workflow was fully documented to ensure experimental reproducibility. In the results that follow, we present relative risk differences, key estimated hyperparameter values, and direct comparisons of model performance.



(A) (B) (C)

FIGURE 5. Modeling results including the poverty covariate. Panel (A) displays observed death counts. Panel (B) shows relative risk maps of the BYM-T model. Panel (C) shows the relative risk differences between the BYM2 model and the BYM-T model.

TABLE 1. Posterior parameter estimates and model comparison for single-covariate studies in West Virginia and Tennessee, with predictive performance assessed using ELPD_{LOO}.

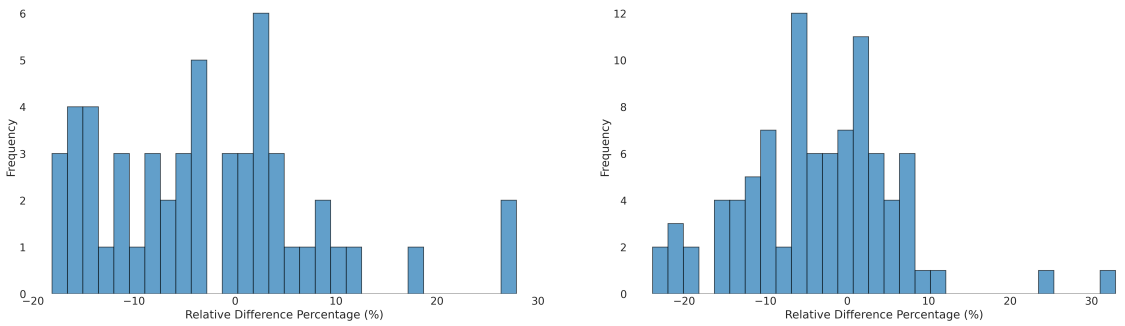
State	Model	$1/\sqrt{\tau_b}$	ϕ	$1-\phi$	$\alpha_1, \alpha_2, \alpha_3$	ELPD_{LOO}	$\text{SE}(\text{ELPD}_{\text{LOO}})$
West Virginia	BYM-T	0.522	-	-	0.201, 0.516, 0.284	-143.230	6.904
	BYM2	0.544	0.502	0.498	-	-146.140	6.553
Tennessee	BYM-T	0.438	-	-	0.421, 0.340, 0.239	-246.736	9.509
	BYM2	0.452	0.275	0.725	-	-247.651	8.157

4.2.2. *Opioid Overdose Risk in West Virginia: Single Covariate Study.* First, we model opioid overdose risk across 53 counties in West Virginia using both the BYM2 and the proposed BYM-T model. In both models, we incorporate poverty as the only covariate. By incorporating poverty as the only covariate, we aim to isolate the impact of including topological summaries of the poverty covariate on disease risk modeling. Specifically, we investigate relative risk differences, proportions of the spatial effects, and model performance. For each county $i = 1, \ldots, 53$, both models define the log relative risk as $\eta_i = \xi + x_i^T \beta + b_i$, where b_i represents the spatial random effect specific to the chosen model (see section 3.4) and x_i^T represents county-level poverty data. Importantly, while the fixed covariate x_i^T uses county-level poverty data, we leverage census tract-level poverty data within each county to calculate the topological summaries of the poverty covariate. Further details on model priors and the experimental framework are provided in section 4.2.1.

Hyperparameter estimates and model comparison criteria are summarized in table 1. The hyperparameter estimates of the BYM2 and BYM-T models are directly comparable due to their appropriate scaling (see section 3.4). To facilitate the comparison, we align the spatial components of the two models: the unstructured component α_1 in the BYM-T model parallels the unstructured component weight $(1-\phi)$ in BYM2, while the structured component α_2 parallels the structured component ϕ in BYM2. The results show a significant redistribution of the weights of spatial components. Specifically, the proportion of weight allocated to the unstructured component decreases from 0.498 (in the BYM2 model) to 0.201 (in the BYM-T model). The majority of this weight is shifted toward the topological component α_3.

To further investigate the impact of the topological summaries on the risk estimates, we compare the relative risk surfaces of the BYM-T and BYM2 models. The fig. 5C shows the numerical differences between the relative risks of two models. The map shows significant changes in risk estimates, with the BYM-T model increasing risk estimates in some counties and decreasing them in others compared to the baseline BYM2 model. To identify the subregional heterogeneities driving these risk differences, we examined census tract-level spatial distributions of poverty values figs. 7 and 8 and summary statistics (table 2) for the counties with the highest relative risk difference percentages (fig. 6A).

Figures 7 and 8 visualize the census-tract level spatial distribution of poverty values for the top four counties with increased and decreased relative difference percentages in relative risk estimates, respectively. Upon comparing these spatial visualizations with statistical summaries (see table 2), a consistent relationship emerges between the census tract-level spatial distributions of poverty values and the resulting risk adjustments. We observe that counties where the BYM-T model predicts increased overdose risk have high-magnitude peaks of census tract-level poverty values



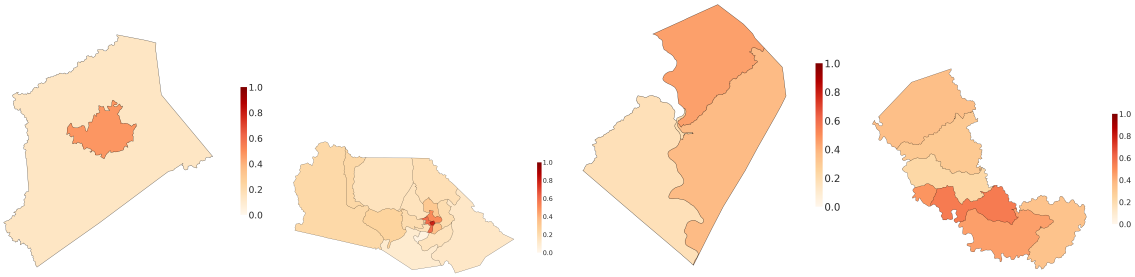
(A) (B)

FIGURE 6. Distribution of percentage relative differences in relative risk for the single covariate study for West Virginia (A) and Tennessee (B).

TABLE 2. Comparison of the top four increases (+) and decreases (-) in relative difference percentages for West Virginia and Tennessee.

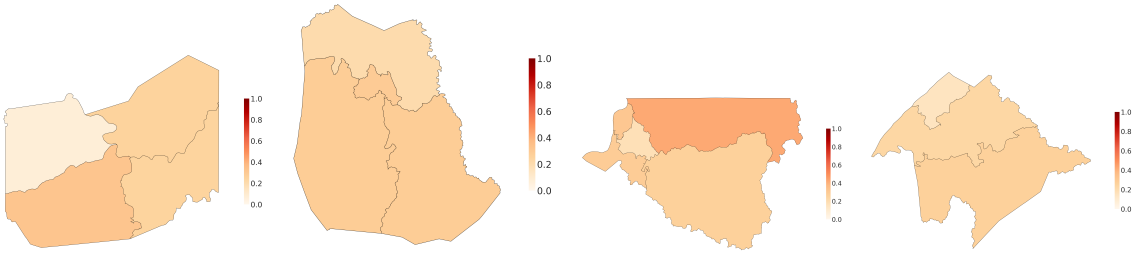
	West Virginia				**Tennessee**			
Trend	FIPS	Diff (%)	Lifespan	Poverty Range	FIPS	Diff (%)	Lifespan	Poverty Range
+	54021	27.875	0.902	0.137–0.473	47183	32.941	3.409	0.143-0.961
	54049	26.606	2.409	0.046–0.788	47063	27.708	1.448	0.088-0.654
	54089	17.238	0.133	0.174–0.446	47117	10.928	0.082	0.108-0.370
	54059	11.409	0.249	0.228–0.559	47087	9.136	0.073	0.190-0.418
−	54091	-16.252	-1.202	0.077–0.326	47069	-20.404	-0.970	0.179-0.299
	54087	-17.405	-0.236	0.205–0.296	47033	-21.801	-0.980	0.204-0.284
	54103	-17.453	-0.953	0.177–0.424	47095	-22.720	-0.932	0.309-0.370
	54095	-18.128	-0.722	0.151–0.264	47169	-23.870	-1.201	0.195-0.214

Note: Trend: (+) = Increase, (−) = Decrease. Diff = Relative Risk Difference.
Lifespan = Scaled Avg. Lifespan. Poverty Range denotes min-max census tract values.



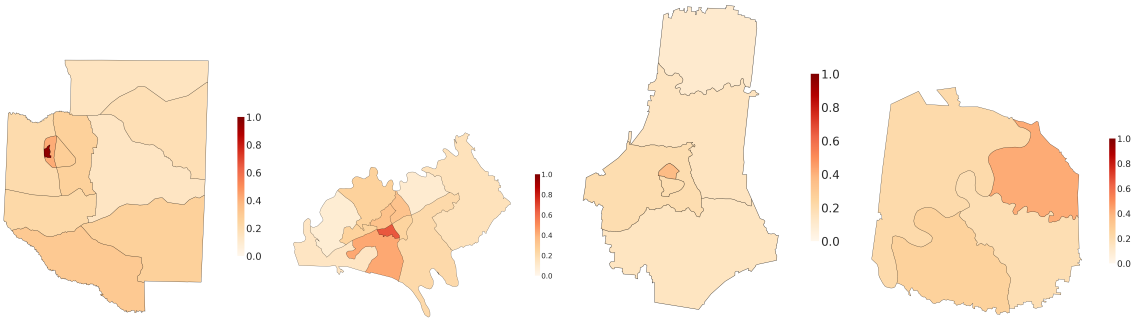
(A) FIPS: 54021 (B) FIPS: 54049 (C) FIPS: 54089 (D) FIPS: 54059

FIGURE 7. Spatial distribution of poverty in the four counties with the largest increases in relative risk difference percentages in West Virginia.



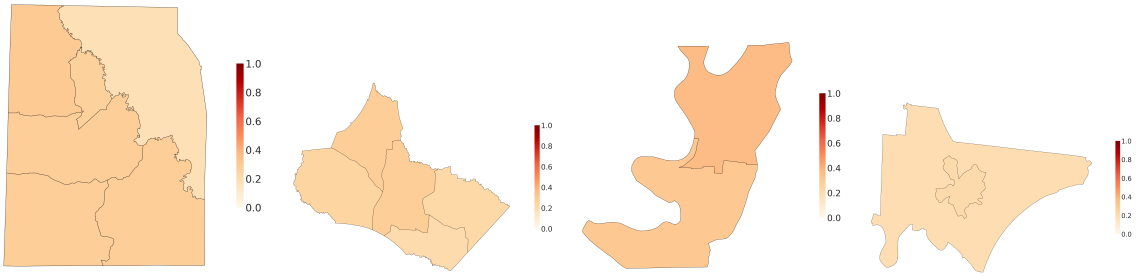
(A) FIPS: 54091 (B) FIPS: 54087 (C) FIPS: 54103 (D) FIPS: 54095

FIGURE 8. Spatial distribution of poverty in the four counties with the largest decrease in relative risk difference percentages in West Virginia.



(A) FIPS: 47183 (B) FIPS: 47063 (C) FIPS: 47117 (D) FIPS: 47087

FIGURE 9. Spatial distribution of poverty in the four counties with the largest increases in relative difference percentages in Tennessee.



(A) FIPS: 47069 (B) FIPS: 47033 (C) FIPS: 47095 (D) FIPS: 47169

FIGURE 10. Spatial distribution of poverty in the four counties with the largest decrease in relative difference percentage in Tennessee.

(also called "high peaks") or high-magnitude census tract-level spatial distribution of poverty values with high scaled topological persistence. Conversely, counties where the overdose risk is adjusted downward tend to exhibit low-magnitude census tract-level spatial distribution of poverty values with low-scaled topological persistence.

For the single-covariate study, we compare the predictive performance of the BYM-T and BYM2 models using ELPD_{LOO}. In terms of model choice, the BYM-T model exhibits a higher ELPD_{LOO} value (-143.230) than the BYM2 model (-146.140); however, the ELPD_{LOO} difference is small relative to its associated standard error (ΔELPD=2.91, SE=2.4), indicating comparable predictive performance.

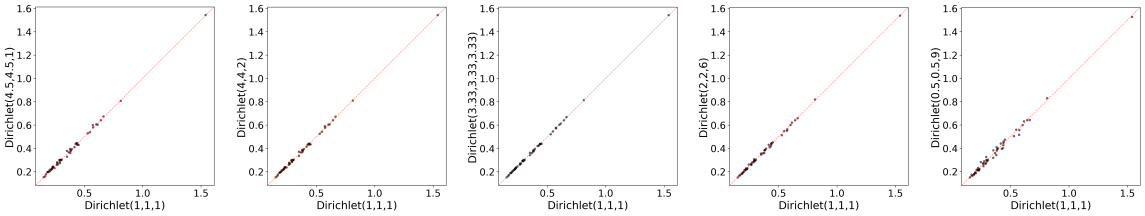


FIGURE 11. Posterior mean relative risk estimates under six prior distributions for the mixing parameters $(\alpha_1, \alpha_2, \alpha_3)$, from the single-covariate BYM-T model adjusted for poverty.

To evaluate how prior selection of the mixing parameters influences model stability, we conducted a sensitivity analysis on the Dirichlet prior for the mixing parameters α_1, α_2, and α_3. We examined five informative priors: $Dirichlet(4.5, 4.5, 1)$, $Dirichlet(4, 4, 2)$, $Dirichlet(3.33, 3.33, 3.33)$, $Dirichlet(2, 2, 6)$, and $Dirichlet(0.5, 0.5, 9)$. These specifications correspond to prior expectations that the topological component explains 10%, 20%, 33.3%, 60%, or 90% of the total spatial random effect, respectively. Additionally, we considered a non-informative $Dirichlet(1, 1, 1)$ prior. This study was carried out independently for two different single-covariate models.

Table 3 shows the resulting posterior means for the model including poverty and the model including unemployment. In both cases, we observe that the overall risk ξ and the total standard deviation $1/\sqrt{\tau_b}$ remain relatively unchanged across all prior specifications. In terms of mixing parameters, we discover that while the structured component α_2 stays relatively constant compared to α_1 and α_3, increasing the prior weight on the topological component α_3 mainly causes a compensatory decrease in the unstructured component α_1. Furthermore, the posterior relative risk estimates show negligible variation (see fig. 11 and fig. 15), indicating that relative risk estimates of the models are robust to the choice of prior.

4.2.3. *Opioid Overdose Risk in Tennessee: Single Covariate Study.* We model opioid overdose risk across 91 counties in Tennessee using both the BYM2 and the proposed BYM-T model. In both models, we incorporate poverty as the only covariate. For each county $i = 1, \ldots, 91$, both models define the log relative risk as $\eta_i = \xi + x_i^T\beta + b_i$, where b_i represents the spatial random effect specific to the chosen model (see section 3.4) and x_i^T represents county-level poverty data. The calculation of topological summaries from census tract-level data and model priors remains consistent with the framework defined in the section 4.2.2

The results show a significant redistribution of the weights of spatial components. Specifically, the proportion of weight allocated to the unstructured component decreases from 0.725 (in the

TABLE 3. Posterior estimates for model parameters under varying Dirichlet prior distributions for the mixing parameters $(\alpha_1, \alpha_2, \alpha_3)$. Results are derived from the single-covariate BYM-T model adjusted for poverty and unemployment.

	Model Adjusted for Poverty				**Model Adjusted for Unemployment**			
Parameter	Mean	SD	3% quant.	97% quant.	Mean	SD	3% quant.	97% quant.
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(1, 1, 1)$								
ξ	-1.051	0.200	-1.416	-0.668	-0.946	0.170	-1.268	-0.630
$1/\sqrt{\tau_b}$	0.522	0.071	0.395	0.657	0.539	0.075	0.403	0.681
α_1	0.201	0.144	0.00	0.457	0.276	0.165	0.000	0.651
α_2	0.516	0.153	0.241	0.804	0.521	0.166	0.206	0.822
α_3	0.284	0.101	0.094	0.471	0.203	0.091	0.034	0.365
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(4.5, 4.5, 1)$								
ξ	-1.078	0.204	-1.468	-0.702	-0.957	0.169	-1.280	-0.648
$1/\sqrt{\tau_b}$	0.518	0.071	0.391	0.656	0.527	0.073	0.395	0.667
α_1	0.338	0.121	0.122	0.567	0.383	0.125	0.153	0.619
α_2	0.467	0.120	0.244	0.688	0.477	0.123	0.252	0.708
α_3	0.195	0.086	0.037	0.353	0.140	0.076	0.012	0.278
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(4, 4, 2)$								
ξ	-1.060	0.204	-1.447	-0.682	-0.954	0.169	-1.266	-0.632
$1/\sqrt{\tau_b}$	0.517	0.070	0.390	0.649	0.529	0.072	0.395	0.661
α_1	0.307	0.118	0.098	0.530	0.351	0.121	0.138	0.583
α_2	0.449	0.118	0.227	0.665	0.462	0.120	0.240	0.687
α_3	0.245	0.088	0.077	0.403	0.187	0.079	0.048	0.335
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(3.33, 3.33, 3.33)$								
ξ	-1.044	0.199	-1.440	-0.694	-0.943	0.166	-1.250	-0.630
$1/\sqrt{\tau_b}$	0.517	0.070	0.389	0.648	0.530	0.071	0.401	0.662
α_1	0.270	0.113	0.066	0.477	0.316	0.117	0.106	0.536
α_2	0.430	0.117	0.204	0.645	0.442	0.120	0.222	0.668
α_3	0.300	0.088	0.137	0.465	0.242	0.081	0.097	0.392
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(2, 2, 6)$								
ξ	-1.018	0.199	-1.395	-0.648	-0.929	0.166	-1.239	-0.615
$1/\sqrt{\tau_b}$	0.513	0.067	0.397	0.644	0.536	0.071	0.410	0.672
α_1	0.202	0.106	0.023	0.396	0.255	0.118	0.042	0.466
α_2	0.403	0.114	0.190	0.615	0.413	0.123	0.178	0.637
α_3	0.395	0.086	0.237	0.555	0.332	0.081	0.184	0.484
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(0.5, 0.5, 9)$								
ξ	-0.999	0.187	-1.352	-0.650	-0.915	0.158	-1.223	-0.626
$1/\sqrt{\tau_b}$	0.507	0.064	0.389	0.627	0.543	0.070	0.415	0.676
α_1	0.095	0.095	0.000	0.274	0.159	0.119	0.000	0.365
α_2	0.422	0.116	0.209	0.642	0.426	0.131	0.180	0.660
α_3	0.483	0.082	0.327	0.633	0.414	0.077	0.264	0.552

BYM2 model) to 0.421 (in the BYM-T model). The majority of this weight is shifted toward the topological component α_3, resulting in refined relative risk adjustments. As shown in fig. 12C, the BYM-T model introduces significant adjustments to risk estimates compared to the BYM2 model, increasing risk in certain counties while decreasing risk in others. To identify the subregional heterogeneities driving these risk differences, we examined census tract-level spatial distributions

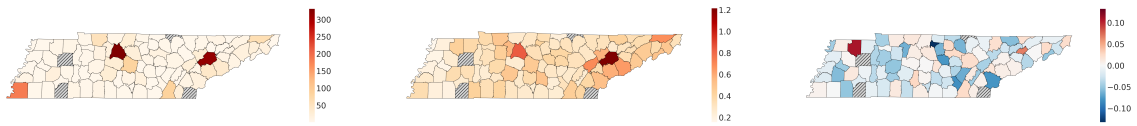
TABLE 4. Posterior parameter estimates and model comparison for multiple-covariate studies in West Virginia, with predictive performance assessed using ELPD_{LOO}.

Model	$\frac{1}{\sqrt{\tau_b}}$	ϕ	1-ϕ	$\alpha_1, \alpha_2, \alpha_3$	ELPD_{LOO}	$\text{SE}(\text{ELPD}_{\text{LOO}})$
BYM-T	0.482	-	-	0.152,0.525,0.323	-142.409	6.898
BYM2	0.467	0.611	0.389	-	-144.576	6.448

of poverty valuesfigs. 9 and 10 and summary statistics (table 2) for the counties with the highest relative risk difference percentages (fig. 6B).

Figures 9 and 10 visualize the census-tract level spatial distribution of poverty values for the top four counties with increased and decreased relative difference percentages in relative risk estimates, respectively. Upon comparing these spatial visualizations with statistical summaries table 2, a consistent relationship emerges between the census tract-level spatial distributions of poverty values and the resulting risk adjustment that aligns with the findings detailed in section 4.2.2. We observe that counties where the BYM-T model predicts increased overdose risk have high peaks of census tract-level poverty values or high-magnitude census tract-level spatial distribution of poverty values with high scaled topological persistence. Conversely, counties where the overdose risk is adjusted downward tend to exhibit low-magnitude census tract-level spatial distribution of poverty values with low-scaled topological persistence.

For the single-covariate study, we compare the predictive performance of the BYM-T and BYM2 models using ELPD_{LOO}. In terms of model choice, the BYM-T model exhibits a higher ELPD_{LOO} value (-246.736) than the BYM2 model (-247.651); however, the ELPD_{LOO} difference is small relative to its associated standard error (ΔELPD=0.91, SE=3.1), indicating comparable predictive performance.



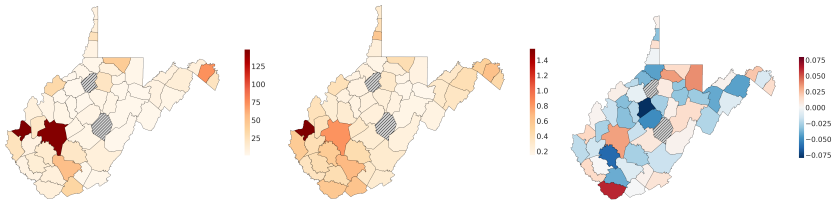
(A) (B) (C)

FIGURE 12. Modeling results including the poverty covariate. Panel (A) displays observed death counts. Panel (B) shows relative risk maps of the BYM-T model. Panel (C) shows the relative risk differences between the BYM2 model and the BYM-T model.

4.2.4. *Opioid Overdose Risk in West Virginia: Multi-Covariate Study.* In the second stage of our analysis, we extend our framework to a multi-covariate setting. In a multi-covariate setting, the topological component represents topological summaries of all the covariates. As covariates, we initially selected 15 candidate variables from the SVI; however, we refined this set through a recursive screening process based on variance inflation factors to address potential multicollinearity (see fig. 13). This process obtained a final set of seven covariates: unemployment, population age above 65, minority population, limited English proficiency, multi-unit housing, household crowding, and group quarter population.

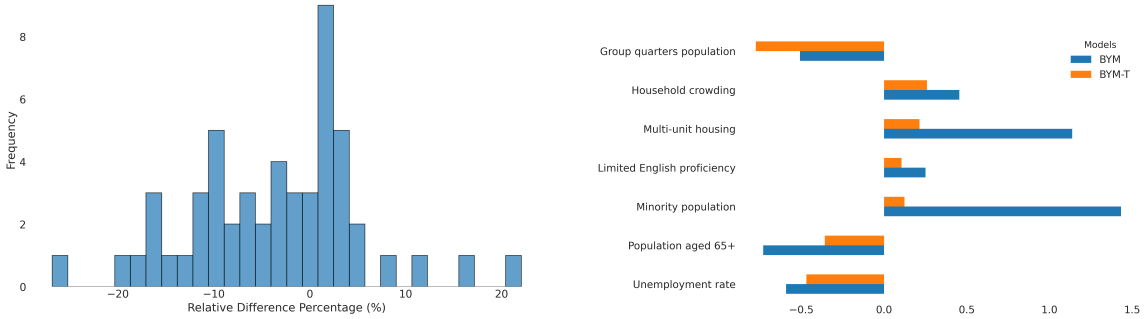
Hyperparameter estimates and model comparison criteria are summarized in table 4. The results show a similar outcome as sections 4.2.2 and 4.2.3 in terms of the distribution of the weights of the spatial components. Specifically, the proportion of weight that is allocated to the unstructured component α_1 decreases from 0.389 (in the BYM2 model) to 0.152 (in the BYM-T model). The majority of this weight is shifted towards the topological component α_3. However, we observed that the inclusion of the topological component leads to a general shrinkage in the magnitude of the fixed covariate coefficients β in the BYM-T model (see fig. 14B).

Regarding the model selection, the BYM-T model exhibits a higher ELPD_{LOO} value (-142.409) than the BYM2 model (-144.576); however, the ELPD_{LOO} difference is small relative to its associated standard error (ΔELPD=2.16, SE=2.1), indicating comparable predictive performance.



(A) (B) (C)

FIGURE 13. All maps present modeling results that include multiple covariates. Panel (A) displays observed deaths. Panel (B) displays the relative risk maps of the BYM-T model. Panel (C) shows the relative risk differences between the BYM2 model and the BYM-T model.



(A) (B)

FIGURE 14. Results from the multi-covariate study. Panel (A) shows the histogram of percentages of relative differences in relative risk. Panel (B) compares the fixed covariate coefficients between the models.

In all the above experiments in section 4.2, no divergences or associated warnings were observed for any parameters. All $\hat{R}$ values were less than 1.01, the ESS values indicated sufficient sampling efficiency, and the rank plots demonstrated good mixing (see appendix A).

5. Discussion and Conclusion

Throughout this paper, we have shown the importance of integrating topological summaries of the subregional-level covariate data into regional-level Bayesian spatial modeling. At the regional level, which has large spatial units, standard Bayesian spatial models are unable to adjust to sharp transitions in risk levels among adjacent neighbors due to over-smoothing of risk estimates. To overcome this limitation, we incorporate topological summaries of the subregional-level covariate data into regional-level models. By providing additional region-specific topological information to regional-level models, we reduce the uncertainty in risk estimates, thereby allowing models to detect sharp transitions in risk among neighbors.

In the BYM2 model, the unstructured component captures everything the model cannot attribute to either fixed covariates or the structured component. Therefore, the unstructured component acts as a residual noise bin. At the regional level, when we introduce the topological component in the BYM-T model, the BYM-T model shifts weight away from the unstructured component and directly into the topological component (e.g., the unstructured component reduces from 0.498 to 0.201 in West Virginia). This shifted weight, which the BYM2 model treated as unexplained localized noise, is, in fact, a topological signal linked to the internal spatial distribution (subregional level) of socioeconomic vulnerability within each region. The BYM2 model lacks the information needed to explain this noise. By providing topological summaries of subregional covariate distributions to the BYM-T model, the BYM-T model gains the information needed to explain what was previously attributed to localized noise, converting it into an informative signal that reflects the internal spatial patterns of socioeconomic vulnerability within each region. By leveraging this explanatory power, our approach reduces the uncertainty of risk estimates, ultimately allowing the model to precisely identify sharp transitions.

Furthermore, the BYM-T model shows a shrinkage effect of the fixed-effects coefficients compared to the BYM2 model. This shrinkage occurs due to collinearity between covariates and spatial random effects [13, 23]. In the proposed BYM-T model, the spatial random effect is decomposed into three parts: a structured component, an unstructured component, and a topological component. However, our results indicate that the introduction of the topological component primarily shifts weight from the unstructured term. Therefore, disentangling whether this coefficient shrinkage is driven by the explicit topological signal or the recalibration of the unstructured noise remains a critical question.

This work is subject to several limitations. Our analysis focused only on connected components, but future extensions could incorporate additional topological features. Moreover, although persistent homology provides robust summaries, choices about filtration methods and covariate selection can influence the results. These choices must be made with careful consideration of the specific application.

By integrating persistent homology into the Bayesian disease-mapping framework, we developed a novel approach that incorporates topological summaries of subregional-level covariate data into the regional-level Bayesian disease-mapping framework. Future work should extend this framework to temporal data, include additional topological features, and apply the method to diverse public health outcomes. Such efforts will further evaluate the generalizability of the BYM-T model and its potential as a practical tool for spatial risk assessment.

Acknowledgments

This initiative is sponsored by the U.S. Department of Veterans Affairs (VA) and utilizes VA-funded resources from the Knowledge Discovery Infrastructure (KDI) at Oak Ridge National Laboratory under the Department of Energy (DOE) Office of Science. The manuscript has been authored by UT-Battelle LLC under contract DE-AC05-00OR22725 with the DOE. The US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce this manuscript or allow others to do so for US government purposes. DOE will provide public access to these results in accordance with the DOE Public Access Plan (`http://energy.gov/downloads/doe-public-access-plan`). The initiative aims to improve care for Veterans within the Veterans Health Administration (VHA) and involves models specifically tailored to the VHA system. The results are not generalizable outside of VHA, and the studies are formally considered non-research by the VHA. Oak Ridge National Laboratory received Institutional Review Board (IRB) approval for the secondary use of patient data, which complies with internal policies and standards. The authors acknowledge the broader partnership and express gratitude to the Veterans receiving care at the VA.

Code Availability

All scripts for data preprocessing, Bayesian spatial modeling, and topological data analysis are openly available at github repository. This includes notebooks that contain the BYM2–T model implementation, simulation framework, and figure reproduction code.

References

[1] O. Abril-Pla, V. Andreani, C. Carroll, L. Dong, C. J. Fonnesbeck, M. Kochurov, R. Kumar, J. Lao, C. C. Luhmann, O. A. Martin, et al., *Pymc: a modern, and comprehensive probabilistic programming framework in python*, PeerJ Computer Science, 9 (2023), p. e1516.

[2] H. Adams and M. Moy, *Topology applied to machine learning: From global to local*, Frontiers in Artificial Intelligence, 4 (2021), p. 668302.

[3] M. A. S. Alhdiri, N. A. Samat, and Z. Mohamed, *Disease mapping for stomach cancer in libya based on besag–york–mollié (bym) model*, Asian Pacific Journal of Cancer Prevention: APJCP, 18 (2017), p. 1479.

[4] S. F. Altekruse, C. M. Cosgrove, W. C. Altekruse, R. A. Jenkins, and C. Blanco, *Socioeconomic risk factors for fatal opioid overdoses in the united states: Findings from the mortality disparities in american communities study (mdac)*, PloS one, 15 (2020), p. e0227966.

[5] V. Y. Au, T. Rosic, N. Sanger, A. Hillmer, C. Chawar, A. Worster, D. C. Marsh, L. Thabane, and Z. Samaan, *Factors associated with opioid overdose during medication-assisted treatment: How can we identify individuals at risk?*, Harm Reduction Journal, 18 (2021), pp. 1–10.

[6] S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical modeling and analysis for spatial data*, Chapman and Hall/CRC, 2003.

[7] A. S. Bennett, H. Guarino, P. C. Britton, D. O'Brien-Mazza, S. H. Cook, F. Taveras, J. Cortez, and L. Elliott, *Us military veterans and the opioid overdose crisis: a review of risk factors and prevention efforts*, Annals of medicine, 54 (2022), pp. 1826–1838.

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

[9] J. Besag, J. York, and A. Mollié, *Bayesian image restoration, with two applications in spatial statistics*, Annals of the institute of statistical mathematics, 43 (1991), pp. 1–20.

[10] R. S. Bivand, E. J. Pebesma, and V. Gomez-Rubio, *Applied spatial data analysis with R*, Springer, 2008.

[11] S. Byers, N. Pritchard, J. Turner, and T. Weighill, *Detecting spatial dependence with persistent homology*, Nonlinear Theory and Its Applications, IEICE, 14 (2023), pp. 106–125.

[12] D. Ciccarone, *The triple wave epidemic: supply and demand drivers of the us opioid overdose crisis*, International Journal of Drug Policy, 71 (2019), pp. 183–188.

[13] D. G. Clayton, L. Bernardinelli, and C. Montomoli, *Spatial correlation in ecological analysis*, International journal of epidemiology, 22 (1993), pp. 1193–1202.

[14] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, *Stability of persistence diagrams*, in Proceedings of the twenty-first annual symposium on Computational geometry, 2005, pp. 263–271.

[15] C. Dean, M. Ugarte, and A. Militino, *Detecting interaction between random region and fixed age effects in disease mapping*, Biometrics, 57 (2001), pp. 197–202.

[16] A. Deas, A. Spannaus, H. Fernando, H. A. Hanson, A. J. Kapadia, J. Trafton, and V. Maroulas, *Identifying spatiotemporal patterns in opioid vulnerability: investigating the links between disability, prescription opioids and opioid-related mortality*, BMC Public Health, 25 (2025), pp. 1–16.

[17] Edelsbrunner, Letscher, and Zomorodian, *Topological persistence and simplification*, Discrete & computational geometry, 28 (2002), pp. 511–533.

[18] H. Edelsbrunner and J. Harer, *Computational topology: an introduction*, American Mathematical Soc., 2010.

[19] M. Feng and M. A. Porter, *Persistent homology of geospatial data: A case study with voting*, SIAM Review, 63 (2021), pp. 67–99.

[20] C. for Disease Control, P. (CDC, and others), *Centers for disease control and prevention/agency for toxic substances and disease registry/geospatial research, analysis, and services program. cdc/atsdr social vulnerability index*, (2018).

[21] A. Hickok, D. Needell, and M. A. Porter, *Analysis of spatial and spatiotemporal anomalies using persistent homology: Case studies with covid-19 data*, SIAM Journal on Mathematics of Data Science, 4 (2022), pp. 1116–1144.

[22] C. D. Hodges, H. M. Stephens, and C. Sedney, *Individual and community characteristics of the opioid crisis in west virginia*, Journal of Drug Issues, 54 (2024), pp. 542–559.

[23] J. S. Hodges and B. J. Reich, *Adding spatially-correlated errors can mess up the fixed effect you love*, The American Statistician, 64 (2010), pp. 325–334.

[24] S. Huber, *Persistent homology in data science*, in International Data Science Conference, Springer, 2020, pp. 81–88.

[25] H. Jalal, J. M. Buchanich, M. S. Roberts, L. C. Balmert, K. Zhang, and D. S. Burke, *Changing dynamics of the drug overdose epidemic in the united states from 1979 through 2016*, Science, 361 (2018), p. eaau1184.
[26] J. E. Kelsall and P. J. Diggle, *Non-parametric estimation of spatial variation in relative risk*, Statistics in medicine, 14 (1995), pp. 2335–2342.
[27] D. Lee and R. Mitchell, *Locally adaptive spatial smoothing using conditional auto-regressive models*, Journal of the Royal Statistical Society Series C: Applied Statistics, 62 (2013), pp. 593–608.
[28] B. G. Leroux, X. Lei, and N. Breslow, *Estimation of disease rates in small areas: a new mixed model for spatial dependence*, in Statistical models in epidemiology, the environment, and clinical trials, Springer, 2000, pp. 179–191.
[29] J. LeSage and R. K. Pace, *Introduction to spatial econometrics*, Chapman and Hall/CRC, 2009.
[30] K. Limbeck and B. Rieck, *Detecting spatial dependence in transcriptomics data using vectorised persistence diagrams*, arXiv preprint arXiv:2409.03575, (2024).
[31] Y. C. MacNab, *Bayesian disease mapping: Past, present, and future*, Spatial Statistics, 50 (2022), p. 100593.
[32] R. Merino, N. Bowden, S. Katamneni, and A. Coustasse, *The opioid epidemic in west virginia*, The Health Care Manager, 38 (2019), pp. 187–195.
[33] H. J. Miller, *Tobler's first law and spatial analysis*, Annals of the association of American geographers, 94 (2004), pp. 284–289.
[34] J. R. Munkres, *Elements of algebraic topology*, CRC press, 2018.
[35] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington, *A roadmap for the computation of persistent homology*, EPJ Data Science, 6 (2017), p. 17.
[36] E. M. Porter, C. T. Franck, and M. A. Ferreira, *Gaussian spatial hierarchical models with icar priors*, in Modeling Spatio-Temporal Data, Chapman and Hall/CRC, 2024, pp. 24–47.
[37] H. Quick, *Estimating county-level mortality rates using highly censored data from cdc wonder*, Preventing chronic disease, 16 (2019), p. E76.
[38] J. N. Rao and I. Molina, *Small area estimation*, John Wiley & Sons, 2015.
[39] A. Riebler, S. H. Sørbye, D. Simpson, and H. Rue, *An intuitive bayesian spatial model for disease mapping that accounts for scaling*, Statistical methods in medical research, 25 (2016), pp. 1145–1165.
[40] N. A. Samat and L. W. Mey, *Malaria disease mapping in malaysia based on besag-york-mollie (bym) model*, in Journal of physics: Conference series, vol. 890, IOP Publishing, 2017, p. 012167.
[41] S. H. Sørbye and H. Rue, *Scaling intrinsic gaussian markov random field priors in spatial modelling*, Spatial Statistics, 8 (2014), pp. 39–51.
[42] A. Vehtari, A. Gelman, and J. Gabry, *Practical bayesian model evaluation using leave-one-out cross-validation and waic*, Statistics and computing, 27 (2017), pp. 1413–1432.
[43] A. Zomorodian and G. Carlsson, *Computing persistent homology*, in Proceedings of the twentieth annual symposium on Computational geometry, 2004, pp. 347–356.

University of Tennessee, Knoxville, Tennessee 37996, USA
Email address, Corresponding author: hashan.dan@gmail.com

TABLE 5. Convergence diagnostics of MCMC sampling for single-covariate BYM2 and BYM-T models

Model	Parameter	$\hat{R}$	ESS Bulk	ESS Tail
BYM2	$\frac{1}{\sqrt{\tau_b}}$	1.00027	6388.0	10163.0
BYM2	ϕ	1.00127	2893.0	5016.0
BYM-T	$\frac{1}{\sqrt{\tau_b}}$	1.00059	4987.0	9293.0
BYM-T	α_1	1.00108	2286.0	4003.0
BYM-T	α_2	1.00063	2572.0	5413.0
BYM-T	α_3	1.00052	6768.0	10667.0

APPENDIX A. SUPPLEMENTARY TABLES AND FIGURES

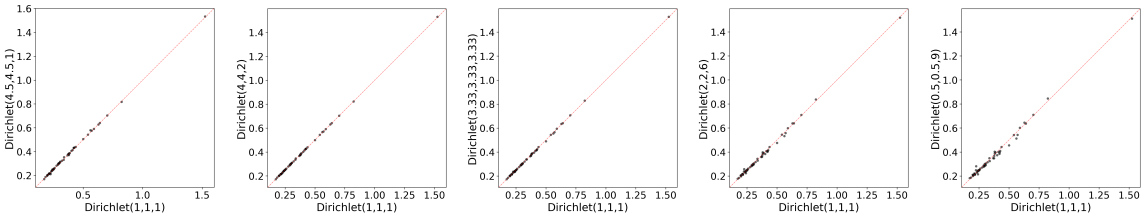
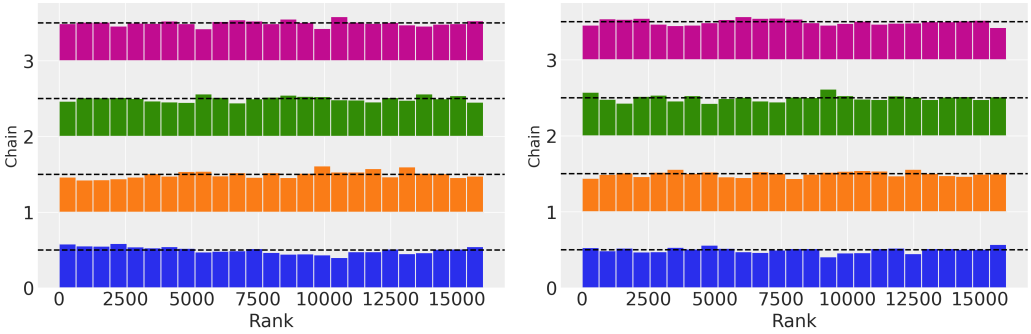
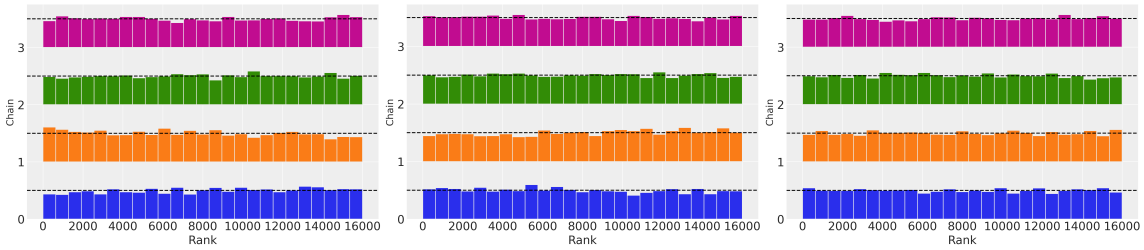


FIGURE 15. Posterior mean relative risk estimates under six prior distributions for the mixing parameters $(\alpha_1, \alpha_2, \alpha_3)$, from the single-covariate BYM-T model adjusted for unemployment.

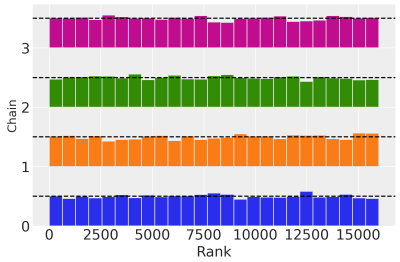


(A) (B)

FIGURE 16. Panels (A) and (B) show rank plots of the ϕ and $1/\sqrt{\tau_b}$, respectively. Results are from the single covariate BYM2 model.



(A)

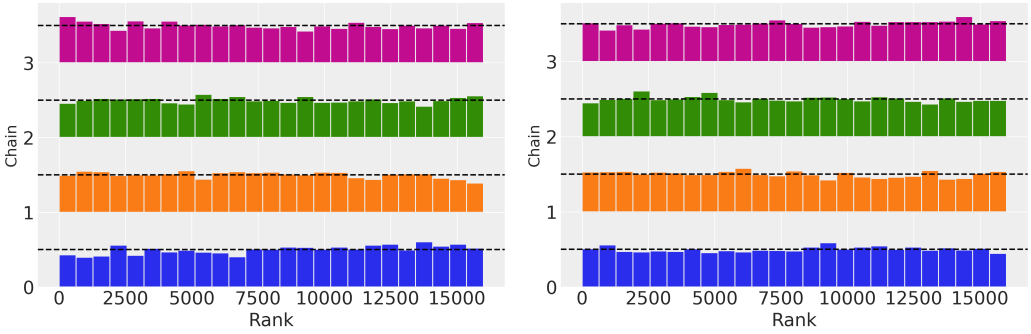


(B)

FIGURE 17. Panel (A) shows rank plots of α_1, α_2, and α_3 (left to right). Panel (B) shows the rank plot of $1/\sqrt{\tau_b}$. Results are from the single covariate BYM-T model.

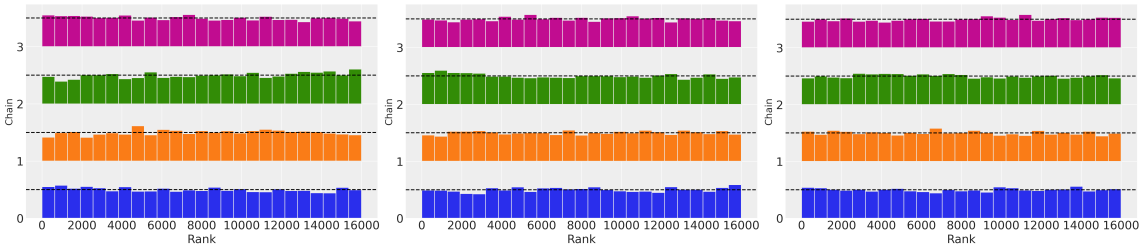
TABLE 6. Convergence diagnostics of MCMC sampling for multi-covariate BYM2 and BYM-T models

Model	Parameter	$\hat{R}$	ESS Bulk	ESS Tail
BYM2	$\frac{1}{\sqrt{\tau_b}}$	1.000315	5262.0	8577.0
BYM2	ϕ	1.001432	3265.0	5983.0
BYM-T	$\frac{1}{\sqrt{\tau_b}}$	1.001451	5164.0	8242.0
BYM-T	α_1	1.002736	3154.0	5486.0
BYM-T	α_2	1.001569	4083.0	6007.0
BYM-T	α_3	1.000057	9930.0	11143.0

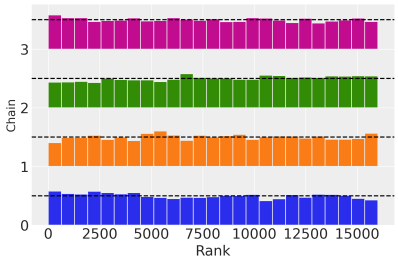


(A) (B)

FIGURE 18. Panels (A) and (B) show rank plots of the ϕ and $1/\sqrt{\tau_b}$, respectively. Results are from the multi-covariate BYM2 model.



(A)



(B)

FIGURE 19. Panel (A) shows rank plots of α_1, α_2, and α_3 (left to right). Panel (B) shows the rank plot of $1/\sqrt{\tau_b}$. Results are from the multi-covariate BYM-T model.

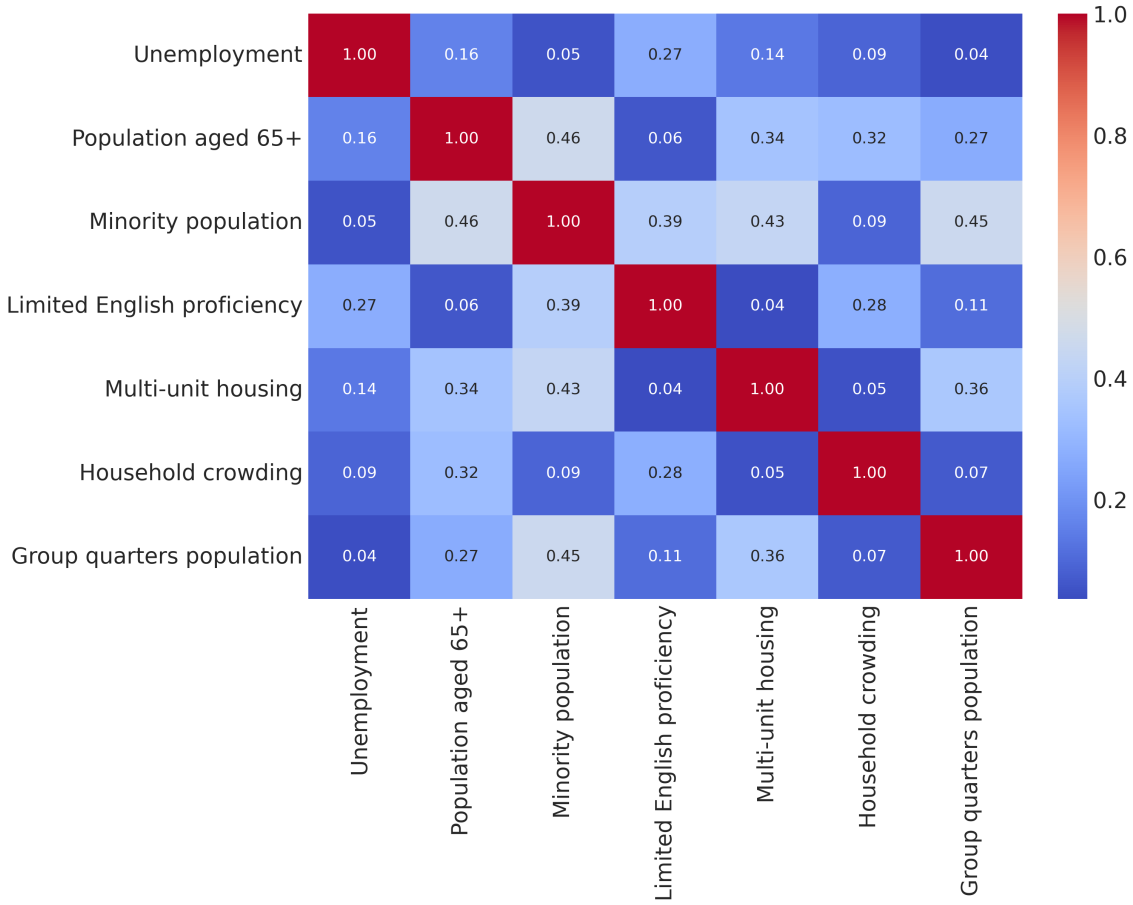


FIGURE 20. Pearson correlation coefficients between selected SVI variables.

Appendix B. Variance Decomposition of the BYM-T Spatial Random Effect

The reparameterized spatial random effect of the BYM-T model is given by

$$b_i^{\text{new}} = \frac{1}{\sqrt{\tau_b}} \left(\sqrt{\alpha_1} v_i + \sqrt{\alpha_2} u_i^* + \sqrt{\alpha_3} h_i \right),$$

where v_i, u_i^*, and h_i are mutually independent, and each is standardized to have a generalized unit variance. Taking the variance of b_i^{new} yields:

$$\begin{aligned}
\text{Var}(b_i^{\text{new}}) &= \text{Var}\left(\frac{1}{\sqrt{\tau_b}} \left(\sqrt{\alpha_1} v_i + \sqrt{\alpha_2} u_i^* + \sqrt{\alpha_3} h_i \right) \right) \\
&= \frac{1}{\tau_b} \left(\alpha_1 \text{Var}(v_i) + \alpha_2 \text{Var}(u_i^*) + \alpha_3 \text{Var}(h_i) \right) \\
&= \frac{1}{\tau_b} \left(\alpha_1 \cdot 1 + \alpha_2 \cdot 1 + \alpha_3 \cdot 1 \right) \\
&= \frac{1}{\tau_b},
\end{aligned}$$

since the mixing parameters are constrained such that $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

This result demonstrates that τ_b retains its strict interpretation as the precision of the total spatial random effect, exactly as formulated in the BYM2 model. Furthermore, because the components are mutually independent and scaled to unit variance, each mixing parameter directly represents the proportion of total spatial variance attributed to its corresponding component: unstructured noise (α_1), spatially structured dependence (α_2), and the topological signal (α_3).

This interpretability is mathematically dependent on the scaling of all three components to unit variance. The BYM2 framework's scaling property is a necessary condition for the BYM-T extension. This makes sure that changes in allocated weight (like from the unstructured component to the topological component) are real reallocations of explanatory power and not just artifacts of different scaling conventions.