Title
Bayesian Hierarchical Models for Multivariate Regionally Aggregated Data using Directed Acyclic Graph Auto-Regressive (DAGAR) models

Author
Gao, Leiwen

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Bayesian Hierarchical Models for Multivariate Regionally Aggregated Data using Directed Acyclic Graph Auto-Regressive (DAGAR) models

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Leiwen Gao

2021

ABSTRACT OF THE DISSERTATION

Bayesian Hierarchical Models for Multivariate Regionally Aggregated Data using Directed Acyclic Graph Auto-Regressive (DAGAR) models

by

Leiwen Gao
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2021
Professor Sudipto Banerjee, Chair

Regional aggregates of health outcomes over delineated administrative units such as counties or zip codes are widely used by epidemiologists to map mortality or incidence rates and better understand geographic variation. Disease mapping is an important statistical tool to assess geographic variation in disease rates and identify lurking environmental risk factors from spatial patterns. Such maps rely upon spatial models for regionally aggregated data, where neighboring regions tend to exhibit more similar outcomes than those farther apart. We contribute to the literature on multivariate disease mapping, which deals with measurements on multiple (two or more) diseases in each region. We aim to disentangle associations among the multiple diseases from spatial autocorrelation in each disease.

We propose two Multivariate Directed Acyclic Graph Autoregression (MDAGAR) models using conditional and joint probability laws respectively to accommodate spatial and inter-disease dependence. The hierarchical construction of conditional MDAGAR imparts flexibility and richness, interpretability of spatial autocorrelation and inter-disease relationships, and computational ease, but depends upon the order in which the diseases are modeled.

To obviate this, we demonstrate how Bayesian model selection and averaging across orders are easily achieved using bridge sampling. We compare our method with a competitor using simulation studies and present an application to multiple cancer mapping using data from the Surveillance, Epidemiology, and End Results (SEER) Program. We also develop a joint MDAGAR model using latent factors, which avoids the disease ordering issue in conditional modelling.

Based on multivariate disease mapping, one often seeks to identify "difference boundaries" that separate adjacent regions with significantly different spatial effects. We adopt a Bayesian multiple-comparison approach for this problem, where we compare all pairs of random effects between neighboring regions. We develop a class of multivariate areally-referenced Dirichlet process (MARDP) models that endow the spatial random effects with a discrete probability law. Within the MARDP framework, the joint MDAGAR model is applied to accommodate spatial and inter-disease dependence for spatial components. We evaluate our method through simulation studies and subsequently present an application to detect difference boundaries for multiple cancers using data from the SEER Program.

The dissertation of Leiwen Gao is approved.

Catherine M. Crespi

Beate Ritz

Donatello Telesca

Sudipto Banerjee, Committee Chair

University of California, Los Angeles

2021

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Dr. Sudipto Banerjee for his guidance, support and encouragement throughout my Ph.D. years at UCLA. His talent, knowledge and deep scientific expertise in mathematics and statistics always inspired me with innovative ideas and thoughts in studies. Under his mentorship, I gradually became a better researcher and developed a better understanding in the field of public health and statistics. I am also grateful for his guidance in my career development, which provided me internship opportunities and helped me determine future direction. I am always very much honored and deeply appreciative to be able to be his student.

I am also very thankful to Dr. Robert Weiss who was the advisor for my master report. He led me into the world of research and showed me how to be a responsible scholar. I appreciate his patience and mentorship, which offered me a good start as a researcher and guided me step by step for my first project. My gratitude also goes to Dr. Antonio Pedro Ramos. I worked as a research assistant for him and learned how to collaborate with other scholars in different fields. Our conversations helped me learn more background knowledge in social science and the importance of applying statistics in real world practice.

I would like to thank Dr. Abhirup Datta for his support. I learned basic model and knowledge of disease mapping from him and was able to move forward to multivariate modeling with his help.

I am especially grateful to my dissertation committee members: Drs. Catherine M. Crespi, Beate Ritz and Donatello Telesca, for their suggestions and comments in my oral defense and dissertation. Communications and discussions with them helped me broaden my views with new ideas and promote my dissertation.

Finally, I am most grateful for the love and support of my family, especially my parents, Yunmin Li and Hongguang Gao, for always being there for me and in favor of my decisions all the time.

The work discussed in this thesis was supported in part by the Division of Mathematical Sciences (DMS) of the National Science Foundation (NSF) under grant 1916349 and by the National Institute of Environmental Health Sciences (NIEHS) under grants R01ES030210 and 5R01ES027027.

VITA

2011–2015 B.S. in Environmental Science, Shandong University

2015–2016 M.S. in Environmental Engineering, University of California, Los Angeles

2016–2018 M.S. in Biostatistics, University of California, Los Angeles

2018–2021 Ph.D. Student and Candidate in Biostatistics, University of California, Los Angeles

PUBLICATIONS

Leiwen Gao, Abhirup Datta, and Sudipto Banerjee (2021). "Hierarchical Multivariate Directed Acyclic Graph Auto-Regressive (MDAGAR) Models for Spatial Diseases Mapping." arXiv preprint arXiv:2102.02911, 2021.

Leiwen Gao, Sudipto Banerjee, and Abhirup Datta (2020). "Spatial Modeling for Correlated Cancers Using Bivariate Directed Graphs." Annals of Cancer Epidemiology, 4(0).

Abhirup Datta, Sudipto Banerjee, James S. Hodges, and **Leiwen Gao** (2019). "Spatial Disease Mapping Using Directed Acyclic Graph Auto-Regressive (DAGAR) Models." Bayesian Analysis, 14(4):1221 - 1244.

CHAPTER 1

Introduction

1.1 Disease mapping for areal data analysis

With increasing interest in analyzing and modelling spatial data, statistical models have been developed to accommodate complex spatial dependencies to generate smoothed maps, estimate model parameters, predict observations at unobserved locations, and test scientific hypotheses [BCG14]. Spatial data can be broadly classified as point-referenced and regionally aggregated (or areal). Point-referenced data are also known as geostatistical data and vary continuously over the domain. By specifying spatial association through structured dependence, models are developed based on a stochastic spatial process and provide spatial prediction (referred to as "kriging") for the point-referenced data setting [BCG14].

Our dissertation focuses on the other type, regionally aggregated or areal data, comprising regional aggregates of health outcomes over delineated administrative units such as states, counties or zip codes. They are widely used by epidemiologists to map counts or rates (e.g., incidence and mortality) and to better understand their geographic variation. Disease mapping, as this exercise is customarily called, employs statistical models to investigate environmental risk factors underlying geographic patterns and present smoothed maps of rates or counts of a disease [Koc05]. Disease maps are used to highlight geographic areas with high and low prevalence, incidence, or mortality rates of cancers, and the variability of such rates over a spatial domain [WCX97]. They can also be used to detect "hot-spots" or spatial clusters which may arise due to common environmental, demographic, or cultural

effects shared by neighboring regions [Ban16]. Maps of crude incidence or mortality rates can be misleading when the population sizes for some of the units are small, which results in large variability in the estimated rates, and makes it difficult to distinguish chance variability from genuine differences. The correct geographic allocation of health care resources can be greatly enhanced by deployment of statistical models that allow a more accurate depiction of true disease rates and their relation to explanatory variables (covariates). Many tasks critical for successful cancer surveillance and control require new inferential methods to handle these complex and often spatially indexed data sets. Since local sample sizes within each spatial region are too low for design-based solutions to attain desired levels of statistical precision [Sch13], much recent work in disease-mapping has been carried out within the context of Bayesian hierarchical models [BCG14]. The body of scientific literature on modern methods for geographic disease mapping is too vast to be reviewed here. Comprehensive reviews of prevalent statistical disease mapping methods and their implementation using available software can be found, among several other sources [BRT05, WC10, WG04, Law13].

For a single disease, there has been a long tradition of employing Markov random fields (MRFs) [RH05] to introduce conditional dependence for the outcome of interest in a region given its neighbors. The outcomes from region units closer to each other are more similar than those recorded in regions farther away. Here, the spatial association across space is constructed through the covariance or precision matrix of the distributions of region-specific latent Gaussian random effects in a hierarchical regression model, which is based on adjacent or neighboring structure of regions. A common approach to build the neighboring structure is to use an undirected graph with the regions constituting the vertices and an edge between two vertices if the corresponding regions share a geographical border as neighbors. Two conspicuous examples are the Conditional Autoregression (CAR) [Bes74, BYM91] and Simultaneous Autoregression (SAR) models [KC08], which build dependence using undirected graphs to model geographic maps.

For a geographic map of the region of interest (e.g., a particular state) delineated by n

distinct administrative regions (e.g., counties or ZIP codes) with clear boundaries separating them, let y_j denote the response related to the disease observed in region j, $j = 1, \ldots, n$, following a Gaussian distribution, i.e. $y_j = \boldsymbol{x}_j^T \boldsymbol{\beta} + w_j + e_j$, where $\boldsymbol{x}_j$ is a $p \times 1$ vector of explanatory covariates, $\boldsymbol{\beta}$ is the slope vector and $e_j \overset{\text{iid}}{\sim} N(0, \tau_e)$ is the random noise with precision τ_e. And $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)^\top$ is a $n \times 1$ vector consisting of spatially associated random effects corresponding to each region j. The CAR model specifies the full conditional distributions with precision τ_j^2,

$$w_j | w_{-j} \sim N(\sum_{j' \neq j} b_{jj'} w_{j'}, \tau_j^2), \tag{1.1}$$

where w_{-j} denotes the vector of observations leaving out the jth one. Through Brook's Lemma [Bro64], the joint distribution of random effects is a multivariate Gaussian with precision matrix $\boldsymbol{D}(\boldsymbol{I} - \boldsymbol{B})$ and written as $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{D}(\boldsymbol{I} - \boldsymbol{B}))$, where $\boldsymbol{B} = \{b_{jj'}\}$ and $\boldsymbol{D}$ is diagonal with τ_j^2 [BCG14]. Let $\boldsymbol{M} = \{m_{jj'}\}$ be the binary adjacency matrix of the geographic map, i.e. $m_{jj'} = 1$ if $j \sim j'$ and 0 otherwise, and $j \sim j'$ indicates regions j and j' are neighbors. By setting $b_{jj'} = m_{jj'}/n_j$ and $\tau_j^2 = \tau n_j$ where n_j is the number of neighbors for region j and τ is the precision scalar, the joint distribution becomes $\boldsymbol{w} \sim N(\boldsymbol{0}, \tau(\boldsymbol{D}_w - \boldsymbol{M}))$ where $\boldsymbol{D}_w$ is diagonal with n_j. However, given $(\boldsymbol{D}_w - \boldsymbol{M})\boldsymbol{1} = \boldsymbol{0}$, the joint distribution of $\boldsymbol{w}$ is improper with singular precision matrix, referred to as the improper CAR (ICAR) model. To resolve this issue, a parameter ρ is added to the model by generalizing the full conditional mean to $E(w_j | w_{-j}) = \rho \sum_{j'} b_{jj'} w_{j'}$, and the joint distribution is redefined as a proper CAR model $\boldsymbol{w} \sim N(\boldsymbol{0}, \tau(\boldsymbol{D}_w - \rho \boldsymbol{M}))$ which is proper for a certain range of ρ. Nevertheless, the interpretation of ρ is difficult since even very high values of ρ may lead to only modest spatial correlation as discussed in section 2.2. The SAR model proceeds by simultaneously modeling the random effects as

$$w_j = \sum_{j' \neq j} b_{jj'} w_{j'} + \epsilon_j, \quad \epsilon_j \overset{\text{ind}}{\sim} N(0, \tau_j^2). \tag{1.2}$$

Equivalently, the joint distribution of random effects can be written as $(\boldsymbol{I}-\boldsymbol{B})\boldsymbol{w}=\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tilde{\boldsymbol{D}})$ and $\tilde{\boldsymbol{D}}$ is the precision matrix with τ_j^2 on the diagonal. Similar to CAR, defining $b_{jj'} = \rho m_{jj'}/n_j$, the joint distribution becomes $\boldsymbol{w} \sim N(\mathbf{0}, (\boldsymbol{I}-\boldsymbol{B})\tilde{\boldsymbol{D}}(\boldsymbol{I}-\boldsymbol{B})^\top)$. Here ρ is called an autoregression parameter and has the similar interpretation issue as the CAR model [Wal04]. More recently, a class of Directed Acyclic Graphical Autoregressive (DAGAR) models have been proposed as a preferred alternative to CAR or SAR models in allowing better identifiability and interpretation of spatial autocorrelation parameters as introduced in Section 2.2.

1.2 Multivariate analysis for disease mapping

For multiple outcomes observed over each unit, multivariate disease mapping is concerned with the analysis of multiple diseases that are associated among themselves as well as across space. It is appropriate when different diseases have been observed over the same spatial units and when the diseases themselves are related to each other, say because they share the same set of genetic and environmental risk factors [LFB17, JCB05, AMA18, SC04]. In other words, we seek models to capture the spatial association for each disease as well as the association between the diseases. When the diseases are inherently related so that the prevalence of one in a region encourages (or inhibits) occurrence of the other on the same unit, there can be substantial inferential benefits in jointly modeling the diseases rather than fitting independent univariate models for each disease [GV03, Mar88, MBB17, Mar13, KB01, CB03, HNF05, DDB08, MMG14]. While it has been assertively demonstrated that independent models for diseases can lead to biased results because of unaccounted associations among the diseases, the current literature is largely based on using CAR models for spatial mapping [KST01, JCB05, JBC07, ZHB09].

Broadly speaking, there are two approaches to multivariate areal modeling based on the CAR model. One approach emerges from hierarchical constructions [JCB05, DZZ06]

where each disease enters the model in a given sequence as conditional structure. For instance, the generalized multivariate CAR (GMCAR) model [JCB05] is discussed in a bivariate case for $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, where $\boldsymbol{w}_i = (w_{i1}, w_{i2}, \ldots, w_{in})^\top$ is the spatial random effect vector for disease i in n regions, $i = 1, 2$. By specifying the marginal distribution of $\boldsymbol{w}_1$, $\boldsymbol{w}_1 \sim N(\boldsymbol{0}, \tau_1(\boldsymbol{D}_w - \rho_1 \boldsymbol{M}))$, and the conditional distribution for $\boldsymbol{w}_2|\boldsymbol{w}_1$, $\boldsymbol{w}_2 = \boldsymbol{A}_{21}\boldsymbol{w}_1 + \boldsymbol{\epsilon}_2$ with $\boldsymbol{\epsilon}_2 \sim N\left(\boldsymbol{0}, \tau_2\left(\boldsymbol{D}_w - \rho_2 \boldsymbol{M}\right)\right)$, the joint distribution for a bivariate spatial process is

$$\begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix} \sim N\left(\boldsymbol{0}, \begin{bmatrix} \tau_1(\boldsymbol{D}_w - \rho_1 \boldsymbol{M}) + \tau_2 \boldsymbol{A}_{21}^\top (\boldsymbol{D}_w - \rho_2 \boldsymbol{M}) \boldsymbol{A}_{21} & \tau_2 \boldsymbol{A}_{21}^\top (\boldsymbol{D}_w - \rho_2 \boldsymbol{M}) \\ \tau_2 (\boldsymbol{D}_w - \rho_2 \boldsymbol{M}) \boldsymbol{A}_{21} & \tau_2 (\boldsymbol{D}_w - \rho_2 \boldsymbol{M}) \end{bmatrix}\right), \tag{1.3}$$

where $\boldsymbol{A}_{21} = \eta_{021}\boldsymbol{I} + \eta_{121}\boldsymbol{M}$ and parameters η_{021} and η_{121} are bridging parameters associating w_{2j} with w_{1j} and $w_{1j'}$, $j \neq j'$, i.e. associating the two different disease in the same region as well as different regions. This bivariate case can be generalized to more diseases. Alternatively, a different class builds upon a linear transformation of latent effects [JCB05, GV03, CB03, Mar13, ZEY05, BHW15a], referred to as joint modelling structure. For example, order-free multivariate CAR (MCAR) models [JBC07] specify the distribution of $\boldsymbol{w} = (\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top, \ldots, \boldsymbol{w}_q^\top)^\top$,

$$\boldsymbol{w}_1 = a_{11}\boldsymbol{f}_1; \quad \boldsymbol{w}_i = a_{i1}\boldsymbol{f}_1 + a_{i2}\boldsymbol{f}_2 + \cdots + a_{ii}\boldsymbol{f}_i,\ i = 2, \ldots, q\ , \tag{1.4}$$

where a_{ih} are coefficient parameters associating random effects for different diseases, $h = 1, \ldots, i$. And $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_i$ are latent effects following CAR models.

While a large amount of multivariate models are developed based upon the CAR models, some spatio-temporal models also utilize Moran's I basis functions for dimension reduction within high-dimensional areal data [BHW15b, BHW18]. For greater interpretability to the spatial autocorrelation parameter, conditional multivariate DAGAR models [GBD20, GDB21] have been developed based on the univariate DAGAR model, which also indicates better interpretation in the association across diseases.

1.3 Difference boundary detection for areal data

Based on disease mapping, one often seeks to identify "difference boundaries" that separate adjacent regions with significantly different spatial effects. The difference boundary is to delineate regions with significantly different spatial random effects from their neighbors. In public health, detecting difference boundaries is useful in exploring significantly different disease mortality and incidence across regions, thus improving decision-making for disease prevention and control, geographic allocation of health care resources, and so on. This exercise has often been referred to as *areal wombling* [Wom51] in spatial data science.

For a single disease, several methods have been developed to detect the difference boundaries on maps. One method that takes on an algorithmic approach is the Boundary Likelihood Values (BLV) [JG03a, JG03b], which fails to adjust for some sources of uncertainty such as the sparsity of data or low population size. Other methods include the Bayesian Information Criterion approach [LBM11] and several model-based frameworks that use hierarchical CAR models such as the LC method [LC05] and site-edge (SE) methods [MCB10] that set priors on edges. For multiple diseases observed in the same spatial unit, a multivariate areal boundary analysis was implemented using a deterministic algorithm to compare posterior estimates from multivariate CAR (MCAR) models [CM07].

The approach we adopt here is to consider this problem as one of Bayesian multiple testing, where we wish to formally evaluate the posterior probability that the spatial random effects from a pair of adjacent regions are different. These posterior probabilities are computed for all pairwise adjacencies on the map and subsequently controlled for (Bayesian) False Discovery Rates (FDR) [MPR04]. In this approach, we will need to ensure that the posterior probability $P(w_{ij} = w_{i'j'} \,|\, \text{Data}, j \sim j')$ is meaningfully defined and can be non-zero, where w_{ij} denotes the spatial effect corresponding to disease i in region j. This means that we must endow the spatial effects with a discrete distribution.

Multivariate spatial models for continuous random effects [Mac18] have demonstrated

the statistical benefits of jointly modeling multiple diseases across areal units. While such models are often constructed using multivariate Markov random fields [Mar88], spatiotemporal models for areal data using Moran's I basis functions for dimension reduction have been developed [BHW15b, BHW18].

For discrete multivariate spatial distributions, one could build upon classes of parametric univariate discrete spatial moving average models (SMA) [BCL12]. However, inference from such models are very sensitive to prior specifications. Instead, we expand upon a demonstrably effective nonparametric approach for univariate boundary detection, i.e. the areally-referenced Dirichlet process (ARDP) model [LBH15, HBL15] and extend them to analyze multiple correlated diseases. More specifically, we achieve probabilistic estimation for difference boundaries by embedding a multivariate areal model within a hierarchical Dirichlet process model. We call this a Multivariate Areal Dirichlet Process (MARDP).

We remark that other authors have adopted different viewpoints on boundary detection from ours. For example, an integrated stochastic process [QBN21] was proposed to infer boundaries based upon continuous gradients as defined in curvilinear wombling [BG06]. While attractive for continuous random fields where a "wombling boundary" is defined as one located in a zone with high directional gradients, our difference boundaries are a subset of administrative boundaries defined on the basis of significantly different spatial effects. A stochastic edge mixed effects (SEME) [GB19] model was used for unknown adjacencies and detected the presence of edges by incorporating covariates. The detection of edges was only used for the improvement of spatial effects estimation but not difference boundary detection.

Turning to FDR-based methods, we note the work for testing an uncountable set of hypothesis tests on Gaussian random fields [PGV04] and the FDR smoothing approach [TKP18] that exploits spatial structure within a multiple-testing problem. The former pertains to point-referenced data, while we focus on areal data. The latter focuses on identifying regions with enriched local fraction of signals against the background, while we intend to ascertain difference boundaries based upon the differences between latent spatial random

effects after accounting for risk factors, confounders and other explanatory variables.

1.4 Contributions and dissertation outline

Directed Acyclic Graphical Autoregressive (DAGAR) models have been proposed to employ directed acyclic graphs as a preferred alternative for univariate disease mapping [DBH19]. A specific motivation for DAGAR models is that they impart greater interpretability to the spatial autocorrelation parameter. In this dissertation, given the advantage of DAGAR models, we develop multivariate DAGAR (MDAGAR) models for multiple disease mapping using the two approaches, i.e. conditional and joint modelling respectively, and help epidemiologists and spatial analysts better interpret the association across diseases. For multivariate difference boundary detection, we extend the ARDP model to a multivariate ARDP (MARDP) framework and utilize the multivariate disease mapping models (MDAGAR and MCAR) to incorporate disease dependence. This multivariate difference boundary analysis framework renders probabilistic estimation for difference boundaries and deliver inference not only for each disease individually but also cross diseases by incorporating association among diseases for spatial components.

The balance of this dissertation proceeds by introducing the DAGAR model for modeling a single disease and different classes of multivariate DAGAR models in Chapter 2. As a starting point of multivariate modelling, a Bivariate DAGAR (BDAGAR) using condition structure is developed for two correlated diseases in practice. The BDAGAR is generalized to conditional multivariate DAGAR models for more than two diseases and a simulation study is conducted to compare with the GMCAR model. In addition, the joint multivariate DAGAR model used for multivariate difference boundary detection is also included. Chapter 3 illustrates the multivariate difference boundary analysis framework MARDP with spatial components constructed by MDAGAR. Under the MARDP framework, we conduct simulation study to demonstrate the robustness of multivariate disease mapping models for

associated diseases. Chapter 4 concludes the dissertation with some discussion.

CHAPTER 2

Multivariate disease mapping using DAGAR model

2.1 Introduction

There is a substantial literature on joint modeling of multiple spatially oriented outcomes, some of which have been cited in Section 1.2. The greater interpretability of DAGAR to the spatial autocorrelation parameter makes it a preferred alternative to CAR or SAR models. While it is possible to model each disease separately using a univariate DAGAR, hence independent of each other, the resulting inference will ignore the association among the diseases. This will be manifested in model assessment because the less dependence among diseases that a model accommodates, the farther away it will be from the joint model in the sense of Kullback-Leibler divergence.

More formally, suppose we have two mutually exclusive sets A and B that contain labels for diseases. Let $\boldsymbol{y}_A$ and $\boldsymbol{y}_B$ be the vectors of spatial outcomes over all regions corresponding to the diseases in set A and set B, respectively. A full joint model $p(\boldsymbol{y})$, where $\boldsymbol{y} = \left(\boldsymbol{y}_A^\top, \boldsymbol{y}_B^\top\right)^\top$, can be written as $p(\boldsymbol{y}) = p(\boldsymbol{y}_A) \times p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_A)$. Let C_1 and C_2 be two nested subsets of diseases in A such that $C_2 \subset C_1 \subset A$. Consider two competing models, $p_1(\boldsymbol{y}) = p(\boldsymbol{y}_A) \times p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_1})$ and $p_2(\boldsymbol{y}) = p(\boldsymbol{y}_A) \times p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_2})$, where $p_1(\cdot)$ and $p_2(\cdot)$ are probability densities constructed from the joint probability measure $p(\cdot)$ by imposing conditional independence such that $p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_A) = p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_1})$ and $p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_A) = p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_2})$, respectively. Both $p_1(\cdot)$ and $p_2(\cdot)$ suppress dependence by shrinking the conditional set A, but $p_2(\cdot)$ suppresses more than $p_1(\cdot)$. We show below that $p_2(\cdot)$ is farther away from $p(\cdot)$ than $p_1(\cdot)$.

A straightforward application of Jensen's inequality yields $\mathbb{E}_{B|C_1}\left[\log \frac{p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_1})}{p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_2})}\right] \geq 0$, where $\mathbb{E}_{B\,|\,C_1}[\cdot]$ denotes the conditional expectation with respect to $p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_1})$. Therefore,

$$\begin{aligned}
\mathrm{KL}(p\|p_2) - \mathrm{KL}(p\|p_2) &= \mathbb{E}_{A,B}\left[\log\left(\frac{p(\boldsymbol{y})}{p_2(\boldsymbol{y})}\right) - \log\left(\frac{p(\boldsymbol{y})}{p_2(\boldsymbol{y})}\right)\right] \\
&= \mathbb{E}_{A,B}\left[\log \frac{p_1(\boldsymbol{y})}{p_2(\boldsymbol{y})}\right] = \mathbb{E}_{A,B}\left[\log \frac{p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_1})}{p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_2})}\right] \\
&= \mathbb{E}_{B,C_1}\left[\log \frac{p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_1})}{p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_2})}\right] = \mathbb{E}_{C_1}\left\{\mathbb{E}_{B\,|\,C_1}\left[\log \frac{p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_1})}{p(\boldsymbol{y}_B \,|\, \boldsymbol{y}_{C_2})}\right]\right\} \geq 0\,.
\end{aligned} \tag{2.1}$$

The equality $\mathbb{E}_{A,B}[\cdot] = \mathbb{E}_{B,C_1}[\cdot]$ in the last row follows from the fact that the argument is a function of diseases in B, C_1 and C_2 and, hence, in B and C_1 because $C_2 \subset C_1$. The argument given in (2.1) is free of distributional assumptions and is linked to the submodularity of entropy and the "information never hurts" principle [CT91, Ban20]. Apart from providing a theoretical argument in favor of joint modeling, (2.1) also notes that models built upon hierarchical dependence structures depend upon the order in which the diseases enter the model. This motivates us to pursue model averaging over the different ordered models in a computationally efficient manner.

Regarding diseases with potential association, the incidence of adenocarcinoma of lung and esophageal cancer have been found to share common risk factors including gastroesophageal reflux disease (GERD), obesity and its associated metabolic syndrome (diabetes, hypertension and hyperlipidemia) [AMA18]. In terms of metabolic mechanisms, it has also been reported that cytochrome P450 2C19 (CYP2C19) may participate in the activation of procarcinogen of both lung and esophageal cancer, and CYP2C19 poor metabolizers (PMs) have higher incidence of two cancers [SC04]. Lung cancer appears to be among the most common second primary cancers in patients with colon cancer [KMW18]. Meanwhile, patients with laryngeal cancer have also been reported to possess high risks of developing second primary lung cancer [ABS10]. We extract the outcome for the incidence of these four can-

cers: lung, esophagus, larynx and colorectal cancer, from the SEER*Stat database using the SEER*Stat statistical software [Nat19]. We start with the analysis of only two cancers (lung and esophageal) in Section 2.3.2 by applying a bivariate DAGAR without too much concern about cancer ordering. Then all four cancers are analyzed using multivariate DAGAR with model selection and averaging to resolve ordering problem for multiple cancers.

The reminder of this chapter is organized as follows. Section 2.2 gives an overview of DAGAR model for single disease mapping. Section 2.3 develops a hierarchical conditional multivariate DAGAR (MDAGAR) model including an application to bivariate case and introduces a bridge sampling method to select the MDAGAR with the best hierarchical order. A simulation study is conducted to compare the MDAGAR with the GMCAR model and illustrates the bridge sampling algorithm's efficacy in selecting the "true" model. At last, the MDAGAR model is applied to the analysis of the incidence of four cancers and discusses different cases with respect to predictors. Section 2.4 introduces another multivariate DAGAR model using joint modelling which solves the ordering issue of multiple diseases and improves computational efficiency.

2.2 Overview of univariate DAGAR model for single disease mapping

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be a graph corresponding to a geographic map, where $\mathcal{V} = \{1, 2, \ldots, n\}$ is a fixed ordering of the vertices of the graph representing clearly delineated regions on the map, and $\mathcal{E} = \{(j, j') : j \sim j'\}$ is the collection of edges between the vertices representing neighboring pairs of regions. The DAGAR model builds a spatial autocorrelation model for a single outcome on $\mathcal{G}$ using an ordered set of vertices in $\mathcal{V}$[DBH19]. Let $N(1)$ be the empty set and let $N(j) = \{j' < j : j' \sim j\}$, where $j \in \mathcal{V} \setminus \{1\}$. Thus, $N(j)$ includes geographic neighbors of region j' that *precede* j in the ordered set $\mathcal{V}$. Let $\{w_i : i \in \mathcal{V}\}$ be a collection of

k random variables defined over the map. DAGAR specifies the following autoregression,

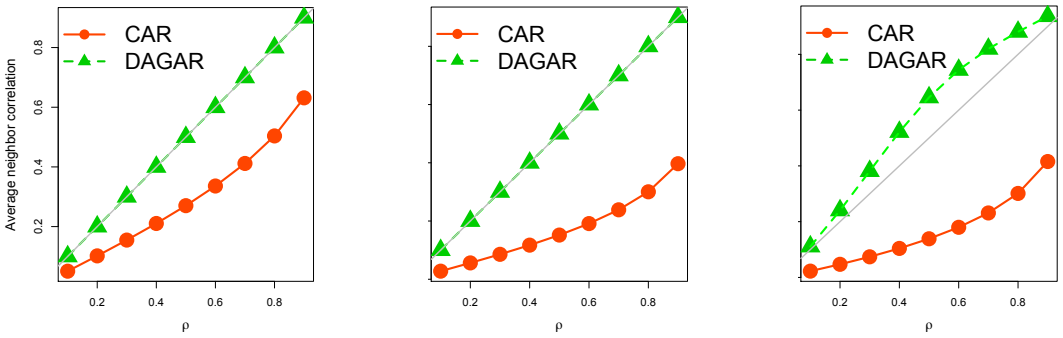
$$w_1 = \epsilon_1; \quad w_j = \sum_{j' \in N(j)} b_{jj'} w_{j'} + \epsilon_j,\ j = 2, \ldots, n\ , \tag{2.2}$$

where $\epsilon_j \overset{\text{ind}}{\sim} N(0, \lambda_j)$ with the precision λ_j, and $b_{jj'} = 0$ if $j' \notin N(j)$. This implies that $\boldsymbol{w} \sim N(\boldsymbol{0}, \tau \boldsymbol{Q}(\rho))$, where $\boldsymbol{Q}(\rho)$ is a spatial precision matrix that depends only upon a spatial autocorrelation parameter ρ and τ is a positive scale parameter. The precision matrix $\boldsymbol{Q}(\rho) = (\boldsymbol{I} - \boldsymbol{B})^\top \boldsymbol{F} (\boldsymbol{I} - \boldsymbol{B})$, $\boldsymbol{B}$ is a $n \times n$ strictly lower-triangular matrix and $\boldsymbol{F}$ is a $n \times n$ diagonal matrix. The elements of $\boldsymbol{B}$ and $\boldsymbol{F}$ are denoted by $b_{jj'}$ and λ_j, respectively, where

$$b_{jj'} = \begin{cases} 0 \ \text{ if } \ j' \notin N(j)\ ; \\ \frac{\rho}{1+(n_{<j}-1)\rho^2} \ \text{ if } \ j = 2, 3, \ldots, n\ ,\ j' \in N(j) \end{cases} \quad \text{and}$$
$$\lambda_j = \frac{1 + (n_{<j} - 1)\rho^2}{1 - \rho^2}\ j = 1, 2 \ldots, n\ , \tag{2.3}$$

$n_{<j}$ is the number of members in $N(j)$ and $n_{<1} = 0$. The above definition of $b_{jj'}$ is consistent with the lower-triangular structure of $\boldsymbol{B}$ because $j' \notin N(j)$ for any $j' \geq j$. The derivation of $\boldsymbol{B}$ and $\boldsymbol{F}$ as functions of a spatial correlation parameter ρ is based upon forming local autoregressive models on embedded spanning trees of subgraphs of $\mathcal{G}$ [DBH19].

The parameter ρ is a measure of spatial correlation between neighboring areas with the value between 0 and 1. Suppose $N(j) = \{j'_1, j'_2, \ldots, j'_m\}$, (2.3) renders a first order auto-regressive (AR(1)) structure for the correlation matrix of $\left(w_j, w_{j'_1}, w_{j'_2}, \ldots, w_{j'_m}\right)$, i.e. $corr\left(w_j, w_{j'_k}\right) = \rho$ and $corr\left(w_{j'_k}, w_{j'_t}\right) = \rho^2$, $k = 1, \ldots, m$, $t = 1, \ldots, m$, $k \neq t$. For $0 \leq \rho < 1$, all λ_i's are positive ensuring a proper probability distribution of $\boldsymbol{w}$ with a positive definite covariance matrix $\boldsymbol{Q}^{-1}(\rho)$, while $\rho = 0$ indicates that all regions are independent. The limiting case of $\rho = 1$ is equivalent to the improper prior with $b_{jj'} = 1/n_{<j}$ and $\lambda_j \propto n_{<j}$, when using the parametrization $\lambda_j = 1 + (n_{<j} - 1)\rho^2$ and absorbing $1/(1-\rho^2)$ into the marginal variance of $\boldsymbol{w}$. Compared to the proper CAR model, DAGAR resolves the issue of a lack



(a) Path graph of length 100 (b) 10×10 grid (c) 48 contiguous US states

Figure 2.1: Average neighbor pair correlations as a function of ρ for proper CAR and DAGAR model. The solid gray line represents $x = y$ line.

of meaningful relationship between ρ and spatial correlation. To illustrate the relationship between ρ and the neighbor-pair correlations for the proper CAR and the DAGAR model, we use two regular graphs and one irregular graph: a simple path graph with 100 vertices which is analogous to a time-series, a two-dimensional 10×10 lattice or grid graph with edges between vertically or horizontally adjacent vertices, and the state map of contiguous United States with 48 states, where two states are said to have an edge if they share a common geographic boundary.

The covariance matrices are generated corresponding to the two models for $\rho \in \{i/10 \mid i = 1, \ldots, 9\}$. Figure 2.1 plots the average neighbor-pair correlation $c(\rho)$ as a function of ρ, where $c(\rho) = \sum_{j' \sim j} cov(w_j, w_{j'})/(2\sqrt{var(w_j)}\sqrt{var(w_{j'})})/(\sum n_j)$, for proper CAR and DAGAR models. For the path and grid graphs, the average neighbor pair correlation $c(\rho)$ for our model is exactly ρ. For the highly irregular United States graph, $c(\rho)$ is much closer to ρ for DAGAR than the proper CAR. For the CAR, even when ρ is close to one, $c(\rho)$ is less than 0.4. In fact, for all three graphs, the average neighbor-pair correlation for the proper CAR model remains modest [DBH19].

2.3 A Conditional Multivariate DAGAR (MDAGAR) model for multiple disease mapping

Modeling multiple diseases will introduce associations among the diseases and spatial dependence for each disease. Let y_{ij} be a disease outcome of interest for disease i in region j. For sake of clarity, we assume that y_{ij} is a continuous variable (e.g., incidence rates) related to a set of explanatory variables through the regression model,

$$y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_i + w_{ij} + e_{ij} \ , \tag{2.4}$$

where $\boldsymbol{x}_{ij}$ is a $p_i \times 1$ vector of explanatory variables specific to disease i within region j, $\boldsymbol{\beta}_i$ are the slopes corresponding to disease i, w_{ij} is a random effect for disease i in region j, and $e_{ij} \overset{ind}{\sim} N(0, (\sigma_i^2)^{-1})$ is the random noise arising from uncontrolled imperfections in the data.

Part of the residual from the explanatory variables is captured by the spatial-temporal effect w_{ij}. Let $\boldsymbol{w}_i = (w_{i1}, w_{i2}, \ldots, w_{in})^\top$ for $i = 1, 2, \ldots, q$. We adopt a hierarchical approach [JCB05], where we specify the joint distribution of $\boldsymbol{w} = (\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top, \ldots, \boldsymbol{w}_q^\top)^\top$ as $p(\boldsymbol{w}) = p(\boldsymbol{w}_1) \prod_{i=2}^{q} p(\boldsymbol{w}_i \,|\, \boldsymbol{w}_{<i})$. We model $p(\boldsymbol{w}_1)$ and each of the conditional densities $p(\boldsymbol{w}_i \,|\, \boldsymbol{w}_{<i})$ with $\boldsymbol{w}_{<i} = (\boldsymbol{w}_1^\top, \ldots, \boldsymbol{w}_{i-1}^\top)^\top$ for $i \geq 2$ as univariate spatial models. The merits of this approach include simplicity and computational efficiency while ensuring that richness in structure is accommodated through the $p(\boldsymbol{w}_i \,|\, \boldsymbol{w}_{<i})$'s. In detail, the multivariate DAGAR (or MDAGAR) model is constructed as

$$\boldsymbol{w}_1 = \boldsymbol{\epsilon}_1; \quad \boldsymbol{w}_i = \boldsymbol{A}_{i1}\boldsymbol{w}_1 + \boldsymbol{A}_{i2}\boldsymbol{w}_2 + \cdots + \boldsymbol{A}_{i,i-1}\boldsymbol{w}_{i-1} + \boldsymbol{\epsilon}_i \ \text{ for } \ i = 2, 3, \ldots, q \ , \tag{2.5}$$

where $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \tau_i \boldsymbol{Q}(\rho_i))$ and $\tau_i \boldsymbol{Q}(\rho_i)$ are univariate DAGAR precision matrices with $\boldsymbol{B}$ and $\boldsymbol{F}$ as in (2.3) with ρ_i. In (2.5), we model $\boldsymbol{w}_1$ as a univariate DAGAR and, progressively, the conditional density of each $\boldsymbol{w}_i$ given $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{i-1}$ is also as a DAGAR for $i = 2, 3, \ldots, q$.

Each disease has its own distribution with its own spatial autocorrelation parameter.

There are q spatial autocorrelation parameters, $\{\rho_1, \rho_2, \ldots, \rho_q\}$, corresponding to the q diseases. This ensures that spatial associations specific to each disease will be captured. Given the differences in the geographic variation of different diseases, this flexibility is desirable. Each matrix $\boldsymbol{A}_{ii'}$ in (2.5) with $i' = 1, \ldots, i-1$ models the association between diseases i and i'. We specify $\boldsymbol{A}_{ii'} = \eta_{0ii'}\boldsymbol{I}_n + \eta_{1ii'}\boldsymbol{M}$, where $\boldsymbol{M}$ is the binary adjacency matrix for the map, i.e., $m_{jj'} = 1$ if $j' \sim j$ and 0 otherwise. Coefficients $\eta_{0ii'}$ and $\eta_{1ii'}$ associate w_{ij} with $w_{i'j}$ and $w_{i'j'}$. In other words, $\eta_{0ii'}$ is the diagonal element in $\boldsymbol{A}_{ii'}$, while $\eta_{1ii'}$ is the element in the j-th row and j'th column if $j' \sim j$. Therefore, for the joint distribution of $\boldsymbol{w}$, if $\boldsymbol{A}$ is the $kq \times kq$ strictly block-lower triangular matrix with (ii')-th block being $\boldsymbol{A}_{ii'} = \boldsymbol{O}$ whenever $i' \geq i$ and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_q^\top)^\top$, then (2.5) renders $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{w} + \boldsymbol{\epsilon}$.

Since $\boldsymbol{I} - \boldsymbol{A}$ is still lower triangular with 1s on the diagonal, it is non-singular with $\det(\boldsymbol{I} - \boldsymbol{A}) = 1$. Writing $\boldsymbol{w} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Lambda})$ and the block diagonal matrix $\boldsymbol{\Lambda}$ has $\tau_1\boldsymbol{Q}(\rho_1), \ldots, \tau_q\boldsymbol{Q}(\rho_q)$ on the diagonal, we obtain $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{Q}_w)$ for $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_q)^\top$ with

$$\boldsymbol{Q}_w = (\boldsymbol{I} - \boldsymbol{A})^\top \boldsymbol{\Lambda} (\boldsymbol{I} - \boldsymbol{A}) \; . \tag{2.6}$$

We say that $\boldsymbol{w}$ follows MDAGAR if $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{Q}_w)$.

Interpretation of $\rho_1, \ldots, \rho_q$ is clear: ρ_1 measures the spatial association for the first disease, while ρ_i, $i \geq 2$, is the residual spatial correlation in the disease i after accounting for the first $i-1$ diseases. Similarly, τ_1 is the spatial precision for the first disease, while τ_i, $i \geq 2$, is the residual spatial precision for disease i after accounting for the first $i-1$ diseases.

We point out two important distinctions from the GMCAR model [JCB05]: (i) instead of using conditional autoregression or CAR for the spatial dependence, we use DAGAR; and (ii) we apply a computationally efficient bridge sampling algorithms [GSM17] to compute the marginal posterior probabilities for each ordered model. The first distinction allows better interpretation of spatial autocorrelation than the CAR models. The second distinction is

of immense practical value and makes this approach feasible for a much larger number of outcomes. Without this distinction, analysts would be dealing with $q!$ models for q diseases and choose among them based upon a model-selection metric. That would be overly burdensome for more than 2 or 3 diseases.

2.3.1 Model Implementation

Let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_q^\top)^\top$ with $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in})^\top$, we extend (2.4) to the following Bayesian hierarchical framework with the posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\tau}, \boldsymbol{\sigma} \,|\, \boldsymbol{y})$ proportional to

$$p(\boldsymbol{\rho}) \times p(\boldsymbol{\eta}) \times \prod_{i=1}^{q} \left\{ IG(1/\tau_i \,|\, a_\tau, b_\tau) \times IG(\sigma_i^2 \,|\, a_\sigma, b_\sigma) \times N(\boldsymbol{\beta}_i \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta^{-1}) \right\}$$
$$\times N(\boldsymbol{w} \,|\, \boldsymbol{0}, \boldsymbol{Q}_w) \times \prod_{i=1}^{q} \prod_{j=1}^{n} N(y_{ij} \,|\, \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_i + w_{ij}, 1/\sigma_i^2) \,, \tag{2.7}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \ldots, \boldsymbol{\beta}_q^\top)^\top$, $\boldsymbol{\tau} = \{\tau_1, \tau_2, \ldots, \tau_q\}$, $\boldsymbol{\sigma} = \{\sigma_1^2, \sigma_2^2, \ldots, \sigma_q^2\}$ and $\boldsymbol{\eta} = \{\boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_q\}$ with $\boldsymbol{\eta}_i = (\boldsymbol{\eta}_{i1}^\top, \boldsymbol{\eta}_{i2}^\top, \ldots, \boldsymbol{\eta}_{i,i-1}^\top)^\top$ and $\boldsymbol{\eta}_{ii'} = (\eta_{0ii'}, \eta_{1ii'})^\top$ for $i = 2, \ldots, q$ and $i' = 1, \ldots, i-1$. For variance parameters $1/\tau_i$ and σ_i^2, $IG(\cdot \,|\, a, b)$ is the inverse-gamma distribution with shape and rate parameters a and b, respectively. For each element in $\boldsymbol{\eta}_i$ we choose a normal prior $N(\mu_{ij}, \sigma_{\eta_{ij}}^2)$, while the prior $N(\boldsymbol{w} \,|\, \boldsymbol{0}, \boldsymbol{Q}_w)$ can also be written as

$$p(\boldsymbol{w}|\boldsymbol{\tau}, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_q, \boldsymbol{\rho}) \propto \tau_1^{\frac{k}{2}} |\boldsymbol{Q}(\rho_1)|^{\frac{1}{2}} \exp\left\{ -\frac{\tau_1}{2} \boldsymbol{w}_1^\top \boldsymbol{Q}(\rho_1) \boldsymbol{w}_1 \right\}$$
$$\times \prod_{i=2}^{q} \tau_i^{\frac{k}{2}} |\boldsymbol{Q}(\rho_i)|^{\frac{1}{2}} \exp\left\{ -\frac{\tau_i}{2} (\boldsymbol{w}_i - \sum_{i'=1}^{i-1} \boldsymbol{A}_{ii'} \boldsymbol{w}_{i'})^\top \boldsymbol{Q}(\rho_i) (\boldsymbol{w}_i - \sum_{i'=1}^{i-1} \boldsymbol{A}_{ii'} \boldsymbol{w}_{i'}) \right\} , \tag{2.8}$$

where $\det(\boldsymbol{Q}(\rho_i)) = \prod_{j=1}^{n} \lambda_{ij}$, and $\boldsymbol{w}_i^T \boldsymbol{Q}(\rho_i) \boldsymbol{w}_i = \lambda_{i1} w_{i1}^2 + \sum_{j=2}^{n} \lambda_{ij} (w_{ij} - \sum_{j' \in N(j)} b_{ijj'} w_{ij'})^2$.

We sample the parameters from the posterior distribution in (2.7) using Markov chain Monte Carlo (MCMC) with Gibbs sampling and random walk metropolis [GL06] as implemented in the `rjags` package within the `R` statistical computing environment. Appendix 2.3.7.1 presents details on the MCMC updating scheme.

2.3.2 An application to bivariate case: Bivariate DAGAR (BDAGAR) model for correlated areal data between two diseases

The MDAGAR in (2.5) can be simplified to a bivariate case by setting $q = 2$ and the hierarchical model becomes

$$p(\boldsymbol{w}_1, \boldsymbol{w}_2) = N(\boldsymbol{w}_1 \,|\, \boldsymbol{0}, \tau_1 \boldsymbol{Q}(\rho_1)) \times N(\boldsymbol{w}_2 \,|\, \boldsymbol{A}_{21}\boldsymbol{w}_1, \tau_2 \boldsymbol{Q}(\rho_2)) \,, \tag{2.9}$$

where $N(\cdot \,|\, \boldsymbol{\mu}, \boldsymbol{Q})$ denotes a normal density with mean μ and precision matrix $\boldsymbol{Q}$. The coefficient matrix $\boldsymbol{A}_{21}$ is simplified to $\boldsymbol{A}_{21} = \eta_0 \boldsymbol{I} + \eta_1 \boldsymbol{M}$. The joint distribution of $\boldsymbol{w} = (\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top)^\top$ is now derived from (2.9) as $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{Q}_w)$, where the precision matrix $\boldsymbol{Q}_w$ is

$$\boldsymbol{Q}_w = \begin{bmatrix} \tau_1 \boldsymbol{Q}(\rho_1) + \tau_2 \boldsymbol{A}_{21}^\top \boldsymbol{Q}(\rho_2) \boldsymbol{A}_{21} & \tau_2 \boldsymbol{A}_{21}^\top \boldsymbol{Q}(\rho_2) \\ \tau_2 \boldsymbol{Q}(\rho_2) \boldsymbol{A}_{21} & \tau_2 \boldsymbol{Q}(\rho_2) \end{bmatrix} \tag{2.10}$$

and the covariance matrix $\boldsymbol{Q}_w^{-1}$ is

$$\boldsymbol{Q}_w^{-1} = \begin{bmatrix} \tau_1^{-1} \boldsymbol{Q}^{-1}(\rho_1) & \tau_1^{-1} \boldsymbol{Q}^{-1}(\rho_1) \boldsymbol{A}_{21}^\top \\ \tau_1^{-1} \boldsymbol{A}_{21} \boldsymbol{Q}^{-1}(\rho_1) & \tau_1^{-1} \boldsymbol{A}_{21} \boldsymbol{Q}^{-1}(\rho_1) \boldsymbol{A}_{21}^\top + \tau_2^{-1} \boldsymbol{Q}^{-1}(\rho_2) \end{bmatrix} . \tag{2.11}$$

We call a normal distribution with the above precision, or covariance, matrix, the BDAGAR model.

The BDAGAR model is applied to analyze a data set including outcomes for 2 cancers, lung (ICD-O-3: C340-C349) and esophagus (ICD-O-3: C150-C159), extracted from the SEER*Stat database using the SEER*Stat statistical software [Nat19]. The outcomes are the 5-year average crude incidence rates per 100,000 population in the years from 2012 to 2016 across 58 counties in California, USA, calculated from the software directly. County-level explanatory variables for each cancer, that possibly affect the incidence rates, are available and include adult cigarette smoking rates in percentage (smoke_{ij}), percentages of residents

younger than 18 years old (young_{ij}), older than 65 years old (old_{ij}), with education level below high school (edu_{ij}) , percentages of unemployed residents (unemp_{ij}), black residents (black_{ij}), male residents (male_{ij}), uninsured residents (uninsure_{ij}), and percentages of families below the poverty threshold (poverty_{ij}). All covariates, except adult cigarette smoking rates, are county attributes extracted from the SEER*Stat database [see] for the years 2012-2016. As a potential common risk factor for both lung and esophageal cancer, adult cigarette smoking rates for 2014-2016 were obtained from the California Tobacco Control Program [Cal18b].

We analyzed this data set using the Bayesian hierarchical model (2.7). The county-level maps of the raw incidence rates per 100,000 population for the two cancers are shown in Figure 2.2. The maps exhibit the evidence of correlation across space and between cancers. Cutoffs for the different levels of incidence rates are quantiles for each cancer. For both lung and esophageal cancer, in general, incidence rates are higher in counties located in the northern areas than those in southern part. The four counties in the center including Amador, Calaveras, Tuolumne and Mariposa have relatively high incidence rates compared to the neighboring counties. Overall, counties with similar levels of incidence rates tend to depict some spatial clustering.

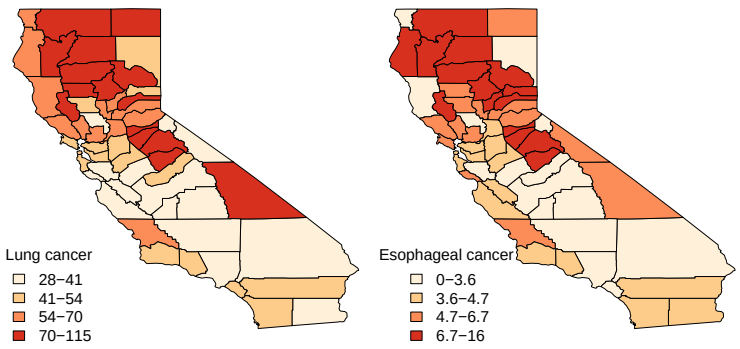


Figure 2.2: Maps of 5-year average crude incidence rates per 100,000 population for lung and esophageal cancer in California, 2012 – 2016.

For our analysis, we specified the following prior distribution,

$$p(\boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\tau}, \boldsymbol{\sigma}, \boldsymbol{w}) = \prod_{i=1}^{2} Unif(\rho_i \,|\, 0, 1) \times \prod_{i=0}^{1} N(\eta_i \,|\, 0, 10^2) \times \prod_{i=1}^{2} N(\boldsymbol{\beta}_i \,|\, 0, 10^3 \boldsymbol{I}) \\ \times \prod_{i=1}^{2} IG(1/\tau_i \,|\, 2, 0.1) \times \prod_{i=1}^{2} IG(\sigma_i^2 \,|\, 2, 1) \times N(\boldsymbol{w} \,|\, 0, \boldsymbol{Q}_w(\tau, \rho)) \,, \qquad (2.12)$$

where $Unif(\cdot \,|\, a, b)$ denotes the Uniform density over $(0, 1)$ and $\boldsymbol{Q}_w(\boldsymbol{\tau}, \boldsymbol{\rho})$ is the BDAGAR precision matrix of $\boldsymbol{w}$ given in [2.10].

We fit the BDAGAR model using the two different cancer orders, i.e. [esophagus] $\times$ [lung $|$ esophagus] and the reverse ordering [lung] $\times$ [esophagus $|$ lung]. We will refer to these orderings simply as [lung $|$ esophagus] and [esophagus $|$ lung], respectively. To compare and assess models, we use the Widely Applicable Information Criterion (WAIC) [Wat10, GHV14] as defined in Section 2.3.7. Table 2.1 presents measures for model fit using the WAIC. We also compare BDAGAR with the "Generalized Multivariate Conditional Autoregression (GMCAR)" models [JCB05]. In both BDAGAR and GMCAR models, the conditional order [esophagus] $\times$ [lung $|$ esophagus] has a smaller WAIC (hence better fit to the data) than the reverse ordering. Meanwhile, within each order, BDAGAR seems to excel over the GMCAR with lower scores in both model fit and effective number of parameters, as seen in the values of $\widehat{elppd}$ and $\hat{p}_{WAIC}$, respectively. The preference of WAIC for [lung $|$ esophagus] is also corroborated by the posterior distribution of η_0 and η_1 from BDAGAR shown in Figure 2.3. In [esophagus $|$ lung], the parameter η_1 has posterior median of -1.94 and a 95% credible interval $(-3.94, -0.58)$. This shows significant negative values that offset part of the significant positive effect of η_0 with a median of 7.58 and a 95% credible interval of $(2.82, 13.94)$. For [lung $|$ esophagus], η_0 is significantly positive with a median of 17.58 and 95% credible interval of $(11.62, 27.84)$, while η_1 tends to be positive with a median of 1.1 but with a 95% credible interval $(-0.77, 2.73)$ that includes 0. Consequently, we present the following results and analysis for [lung $|$ esophagus] which seems to be the preferred model.

Table 2.1: Model comparison using WAIC statistics for cancer data analysis.

Model	lppd	p_{WAIC}	WAIC
BDAGAR (esophagus | lung)	-261.31	45.32	613.27
BDAGAR (lung | esophagus)	-155.12	51.72	413.68
GMCAR (esophagus | lung)	-264.51	46.09	621.19
GMCAR (lung | esophagus)	-156.51	52.05	417.12

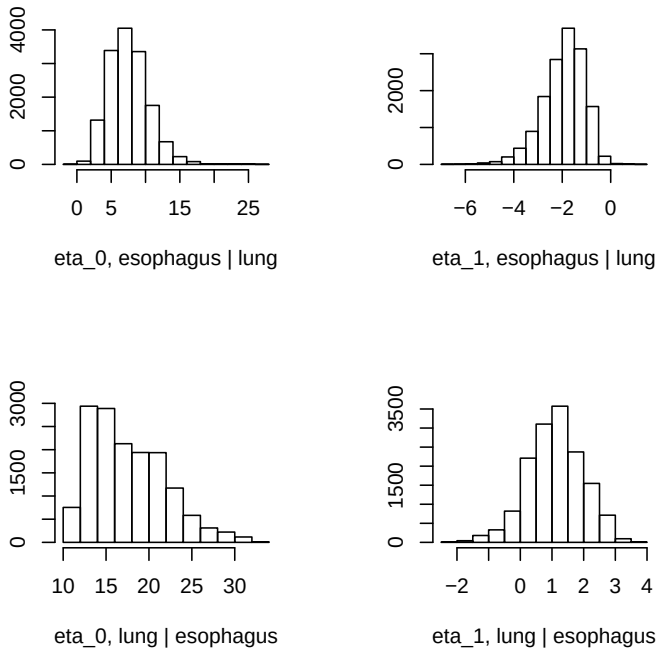


Figure 2.3: Posterior samples of linking parameters η_0, η_1 from BDAGAR model.

Table 2.2 summarizes the parameter estimates from the BDAGAR model corresponding to [lung | esophagus]. For fixed effects, the increasing percentage of residents younger than 18 years old significantly reduces the incidence rate for esophageal cancer, while the percentage of residents older than 65 years old has a significantly opposite effect for lung cancer. Unsurprisingly, higher adult cigarette smoking rates significantly increase the incidence rates for both lung and esophageal cancer. After accounting for these explanatory variables, the residual random effects still exhibit spatial association patterns for both cancers. Turning to spatial correlations, ρ_1 measures the residual spatial correlation (posterior mean 0.08) for

esophageal cancer after accounting for the explanatory variables and ρ_2 measures the spatial correlation (posterior mean 0.5) for lung cancer after accounting for the explanatory variables and also the effect of esophageal cancer. The small point estimates and narrower credible interval for ρ_1 indicate greater confidence in weaker spatial correlation for esophageal cancer; the moderate value of ρ_2 and a wider credible interval suggest higher spatial correlation for lung cancer. Turning to the spatial precision of random effects for each cancer, the estimates of $\{\tau_1, \tau_2\}$ are indicative of esophageal cancer having larger variability, although we must keep in mind that τ_2 is the conditional marginal precision for lung cancer after accounting for esophageal cancer and, therefore, may not be directly comparable to τ_1.

Table 2.2: Parameter estimates (posterior means) for the California cancer incidence rate data from BDAGAR model. Numbers inside braces indicate the lower and upper bounds for the 95% credible intervals.

Parameters	Esophagus cancer	Lung cancer
intercept	18.75 (4.55, 32.72)	7.19 (-47.07, 61.87)
smoke	0.27 (0.12, 0.41)	1.27 (0.28, 2.3)
young	-0.23 (-0.45, -0.01)	-0.75 (-1.94, 0.44)
old	0.14 (-0.03, 0.31)	2.61 (1.62, 3.61)
edu	0.02 (-0.1, 0.14)	-0.25 (-1.04, 0.54)
unemp	-0.07 (-0.26, 0.12)	0.52 (-0.79, 1.84)
black	0.16 (-0.08, 0.39)	0.8 (-0.82, 2.41)
male	-0.04 (-0.19, 0.12)	0.14 (-0.95, 1.26)
uninsure	-0.31 (-0.53, -0.09)	-0.08 (-1.11, 0.94)
poverty	0.32 (-0.33, 0.96)	0.23 (-3.96, 4.48)
ρ_i	0.08 (0, 0.25)	0.5 (0.03, 0.97)
τ_i	2.72 (0.96, 6.69)	19.41 (2.47, 54.36)
σ^2_{ei}	2.05 (1.39, 3.05)	0.93 (0.18, 3.87)

Figure 2.4 shows the estimated correlation between lung and esophageal cancer in each of 58 counties. This map also seems to be consistent with the estimates of η. Correlations between lung and esophageal cancers in all counties are significantly positive with large means at around $0.97 - 1$ which are due to the highly positive values in η_0. This indicates that esophageal cancer is highly correlated with lung cancer. However, in general, the correlation between the two cancers increases slightly from the center to marginal areas,

especially for those with fewer counties in the neighborhood. Finally, Figure 2.5 provides

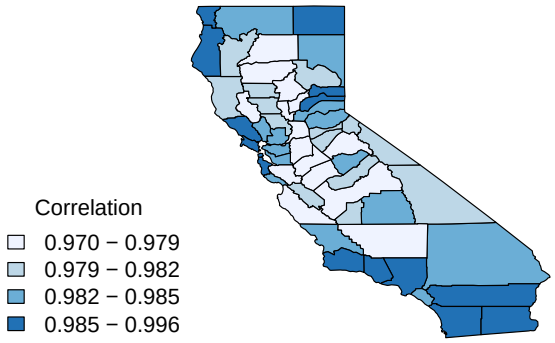


Figure 2.4: Estimated correlation between lung and esophagus cancer in each of 58 counties of California.

further visual corroboration of the goodness of fit for the BDAGAR mode corresponding to [lung | esophagus]. Here, we see that the posterior mean of the incidence rates for lung and esophageal cancer are very consistent with the raw incidence rates shown in Figure 2.2. Given the significant effect of adult cigarette smoking rates on incidence rates for both cancers, the higher fitted incidence rates in the northern areas are in accordance with higher smoking rates in same counties as shown in Figure 2.12. Though the smoking rates are also high in the middle part, the relatively lower fitted incidence rates may be due to the offset of negative spatial random effects for these counties.

2.3.3 Model Selection via Bridge Sampling

It is clear from (2.5) that each ordering of diseases in MDAGAR will produce a different model in terms of the conditional specifications. For the above bivariate case, it is convenient to compare only two models (orders) by the significance of parameter estimates as well as model performance. However, when there are more than two diseases involved in the model,

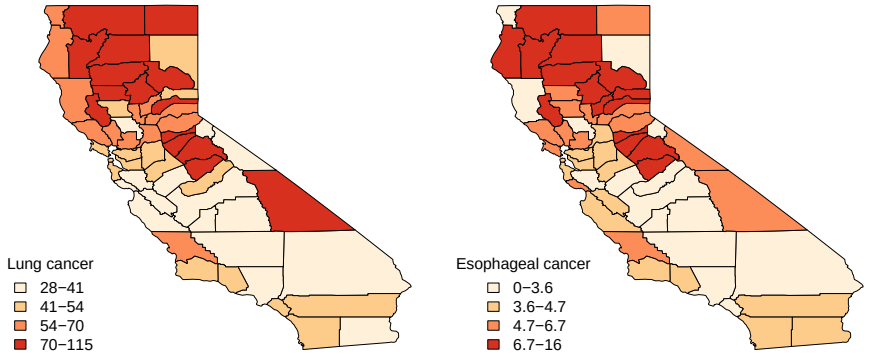


Figure 2.5: Maps of posterior mean incidence rates per 100,000 population for lung and esophagus cancer in California.

at least six models (for three diseases) will be fitted and comparing all models become cumbersome or even impracticable.

Instead, we pursue model averaging of MDAGAR models. Given a set of $T = q!$ candidate models, say $M_1, \ldots, M_T$, Bayesian model selection and model averaging calculates

$$p(M = M_t|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|M = M_t)p(M = M_t)}{\sum_{j=1}^{T} p(\boldsymbol{y}|M = M_j)p(M = M_j)}, \tag{2.13}$$

for $t = 1, \ldots, T$ [HMR99]. Computing the marginal likelihood $p(\boldsymbol{y} \mid M_t)$ in (2.13) is challenging. Methods such as importance sampling [PNT14] and generalized harmonic mean [GD94] have been proposed as stable estimators with finite variance, but finding the required importance density with strong constraints on the tail behavior relative to the posterior distribution is often challenging. Bridge sampling estimates the marginal likelihood (i.e. the normalizing constant) by combining samples from two distributions: a bridge function $h(\cdot)$ and a proposal distribution $g(\cdot)$ [GSW17]. Let $\boldsymbol{\theta}_t = \{\boldsymbol{\beta}_t, \boldsymbol{\sigma}_t, \boldsymbol{\rho}_t, \boldsymbol{\tau}_t, \boldsymbol{\eta}_{2,t}, \ldots, \boldsymbol{\eta}_{q,t}\}$ be the set of parameters in model M_t with prior $p(\boldsymbol{\theta}_t \mid M_t)$ as defined in the first row of (2.7). Based on

the identity,

$$1 = \frac{\int p(\boldsymbol{y}|\boldsymbol{\theta}_t, M_t)p(\boldsymbol{\theta}_t|M_t)h(\boldsymbol{\theta}_t|M_t)g(\boldsymbol{\theta}_t|M_t)d\boldsymbol{\theta}_t}{\int p(\boldsymbol{y}|\boldsymbol{\theta}_t, M_t)p(\boldsymbol{\theta}_t|M_t)h(\boldsymbol{\theta}_t|M_t)g(\boldsymbol{\theta}_t|M_t)d\boldsymbol{\theta}_t},$$

a current version of the bridge sampling estimator is

$$\begin{aligned} p(\boldsymbol{y}|M = M_t) &= \frac{E_{g(\boldsymbol{\theta}_t|M_t)}[p(\boldsymbol{y}|\boldsymbol{\theta}_t, M_t)p(\boldsymbol{\theta}_t|M_t)h(\boldsymbol{\theta}_t|M_t)]}{E_{p(\boldsymbol{\theta}_t|\boldsymbol{y},M_t)}[h(\boldsymbol{\theta}_t|M_t)g(\boldsymbol{\theta}_t|M_t)]} \\ &\approx \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} p(\boldsymbol{y}|\tilde{\boldsymbol{\theta}}_{t,i}, M_t)p(\tilde{\boldsymbol{\theta}}_{t,i}|M_t)h(\tilde{\boldsymbol{\theta}}_{t,i}|M_t)}{\frac{1}{N_1}\sum_{j=1}^{N_1} h(\boldsymbol{\theta}^{\star}_{t,j}|M_t)g(\boldsymbol{\theta}^{\star}_{t,j}|M_t)} \end{aligned} \tag{2.14}$$

where $\boldsymbol{\theta}^{\star}_{t,j} \sim p(\boldsymbol{\theta}_t \,|\, \boldsymbol{y}, M_t), j = 1, \ldots, N_1$, are N_1 posterior samples and $\tilde{\boldsymbol{\theta}}_{t,i} \sim g(\boldsymbol{\theta}_t|M_t), i = 1, \ldots, N_2$, are N_2 samples drawn from the proposal distribution [GSM17]. The likelihood $p(\boldsymbol{y} \,|\, \boldsymbol{\theta}_t, M = M_t)$ is obtained by integrating out $\boldsymbol{w}$ from (2.7) as

$$N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}, \left[\boldsymbol{Q}_w^{-1}(\boldsymbol{\rho}_t, \boldsymbol{\tau}_t, \boldsymbol{\eta}_{2,t}, \ldots, \boldsymbol{\eta}_{q,t}) + diag(\boldsymbol{\sigma}_t) \otimes \boldsymbol{I}_k\right]^{-1}), \tag{2.15}$$

given that $diag(\boldsymbol{\sigma})$ is a diagonal matrix with $\sigma_i^2, i = 1, \ldots, q$, on the diagonal, and $\boldsymbol{X}$ is the design matrix with $\boldsymbol{X}_i$ as block diagonal where $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ik})^{\top}$. The bridge function $h(\boldsymbol{\theta}_t|M_t)$ is specified by the optimal choice [MW96],

$$h(\boldsymbol{\theta}_t|M_t) = C\frac{1}{s_1 p(\boldsymbol{y}|\boldsymbol{\theta}_t, M_t)p(\boldsymbol{\theta}_t|M_t) + s_2 p(\boldsymbol{y}|M_t)g(\boldsymbol{\theta}_t|M_t)} \tag{2.16}$$

where C is a constant. Inserting (2.16) in (2.14) yields the estimate of $p(\boldsymbol{y}|M = M_t)$ after convergence of an iterative scheme [MW96] as

$$\hat{p}(\boldsymbol{y}|M_t)^{(t+1)} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 \hat{p}(\boldsymbol{y}|M_t)^{(t)}}}{\frac{1}{N_1}\sum_{j=1}^{N_1} \frac{1}{s_1 l_{1,j} + s_2 \hat{p}(\boldsymbol{y}|M_t)^{(t)}}} \tag{2.17}$$

where $l_{1,j} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{\star}_{t,j}, M_t)p(\boldsymbol{\theta}^{\star}_{t,j}|M_t)}{g(\boldsymbol{\theta}^{\star}_{t,j}|M_t)}$, $l_{2,i} = \frac{p(\boldsymbol{y}|\tilde{\boldsymbol{\theta}}_{t,i}, M_t)p(\tilde{\boldsymbol{\theta}}_{t,i}|M_t)}{g(\tilde{\boldsymbol{\theta}}_{t,i}|M_t)}$, $s_1 = \frac{N_1}{N_1+N_2}$ and $s_2 = \frac{N_2}{N_1+N_2}$.

Given the log marginal likelihood estimates from `bridgesampling`, the posterior model probability for each model is calculated from (2.13) by setting prior probability of each model $p(M = M_t)$. For Bayesian model averaging (BMA), the model averaged posterior distribution of a quantity of interest Δ is obtained as $p(\Delta \,|\, \boldsymbol{y}) = \sum_{t=1}^{T} p(\Delta \,|\, M = M_t, \boldsymbol{y}) p(M = M_t \,|\, \boldsymbol{y})$ [HMR99], and the posterior mean is

$$E(\Delta \,|\, \boldsymbol{y}) = \sum_{t=1}^{T} E(\Delta \,|\, M = M_t, \boldsymbol{y}) p(M = M_t \,|\, \boldsymbol{y}) \,. \tag{2.18}$$

Setting $\Delta = \{\boldsymbol{\beta}, \boldsymbol{w}\}$ fetches us the model averaged posterior estimates for spatial random effects as well as calculating the posterior mean incidence rates as discussed in Section 2.3.5.

2.3.4 Simulation

We simulated two different experiments to evaluate the performance of MDAGAR model and model selection. The first experiment was designed to evaluate MDAGAR's inferential performance against GMCAR. The second experiment aimed to ascertain the effectiveness of the bridge sampling algorithm (Section 2.3.3) in preferring models with a correct "ordering" of the diseases in the model.

2.3.4.1 Data generation

We compare MDAGAR's inferential performance with GMCAR [JCB05]. We chose the 48 states of the contiguous United States as our underlying map, where two states are treated as neighbors if they share a common geographic boundary. We generated our outcomes y_{ij} using the model in (2.4) with $q = 2$, i.e., two outcomes, and two covariates, $\boldsymbol{x}_{1j}$ and $\boldsymbol{x}_{2j}$, with $p_1 = 2$ and $p_2 = 3$. We fixed the values of the covariates after generating them from $N(\boldsymbol{0}, \boldsymbol{I}_{p_i})$, $i = 1, 2$, independent across regions. The regression slopes were set to $\boldsymbol{\beta}_1 = (1, 5)^\top$ and $\boldsymbol{\beta}_2 = (2, 4, 5)^\top$.

Turning to the spatial random effects, we generated values of $\boldsymbol{w} = \left(\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top\right)^\top$ from a $N(\boldsymbol{0}, \boldsymbol{Q}_w)$ distribution with the precision matrix defined in (2.10). We set $\tau_1 = \tau_2 = 0.25$, $\rho_1 = 0.2$ and $\rho_2 = 0.8$ and take $\boldsymbol{Q}(\rho_i) = \boldsymbol{D}(\rho_i)^{-1}$, where $\boldsymbol{D}(\rho_i) = \exp(-\phi_i d(j, j'))$, $\phi_i = -\log(\rho_i)$ is the spatial decay for disease i and $d(j, j')$ refers to the distance between the embedding of the jth and j'th vertex. The vertices are embedded on the Euclidean plane and the centroid of each state is used to create the distance matrix. Using this exponential covariance matrix to generate the data offers a "neutral" ground to compare the performance of MDAGAR with GMCAR. We specified $\boldsymbol{A}_{12}$ using fixed values of $\boldsymbol{\eta} = \{\eta_{021}, \eta_{121}\}$. Here, we considered three sets of values for $\boldsymbol{\eta}$ to correspond to low, medium and high correlation among diseases. We fixed $\boldsymbol{\eta} = \{0.05, 0.1\}$ to ensure an average correlation of 0.15 (range 0.072 - 0.31); $\boldsymbol{\eta} = \{0.5, 0.3\}$ with an average correlation of 0.55 (range 0.45 - 0.74); and $\boldsymbol{\eta} = \{2.5, 0.5\}$ with a mean correlation of 0.89 (range 0.84 - 0.94). We generated w_{ij}'s for each of the above specifications for $\boldsymbol{\eta}$ and, with the values of w_{ij} generated as above, we generated the outcome $y_{ij} \sim N(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_i + w_{ij}, 1/\sigma_i^2)$, where $\sigma_1^2 = \sigma_2^2 = 0.4$. We repeated the above procedure to replicate 85 data sets for each of the three specifications of $\boldsymbol{\eta}$.

For our second experiment, we generated a data set with $q = 3$ diseases. We extended the above setup to include one more disease. We generated y_{ij}'s from (2.4) with the value of $\boldsymbol{x}_{3j}$ fixed after being generated from $N(\boldsymbol{0}, \boldsymbol{I}_3)$, $\boldsymbol{\beta}_3 = (5, 3, 6)^\top$ and $\sigma_3^2 = 0.4$. Let $[i, j, k]$ denote the model $p(\boldsymbol{w}_i) \times p(\boldsymbol{w}_j \,|\, \boldsymbol{w}_i) \times p(\boldsymbol{w}_k \,|\, \boldsymbol{w}_j, \boldsymbol{w}_i)$. For three diseases the six resulting models are denoted as $M_1 = [1, 2, 3]$, $M_2 = [1, 3, 2]$, $M_3 = [2, 1, 3]$, $M_4 = [2, 3, 1]$, $M_5 = [3, 1, 2]$ and $M_6 = [3, 2, 1]$.

Each of the six models imply a corresponding joint distribution $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{Q}_w)$ which is used to generate the w_{ij}'s. Let the parenthesized suffix (i) denote the disease in the ith order. For example, in $M_2 = [1, 3, 2]$, we write $\boldsymbol{w}$ in the form of (2.5) as

$$\boldsymbol{w}_1 \sim \epsilon_{(1)}; \quad \boldsymbol{w}_3 = \boldsymbol{A}_{(21)} \boldsymbol{w}_1 + \epsilon_{(2)}; \quad \boldsymbol{w}_2 = \boldsymbol{A}_{(31)} \boldsymbol{w}_1 + \boldsymbol{A}_{(32)} \boldsymbol{w}_3 + \epsilon_{(3)} \ ,$$

where $\epsilon_{(i)} \sim N(\mathbf{0}, \tau_{(i)}\boldsymbol{Q}\left(\rho_{(i)}\right))$ with $\boldsymbol{Q}(\rho_{(i)}) = \boldsymbol{D}(\rho_{(i)})^{-1}$ as in the first experiment, and $\boldsymbol{A}_{(ii')} = \eta_{0(ii')}\boldsymbol{I} + \eta_{1(ii')}\boldsymbol{M}$ is the coefficient matrix associating random effects for diseases in the ith and i'th order. We set $\tau_{(1)} = \tau_{(2)} = \tau_{(3)} = 0.25$, $\rho_{(1)} = 0.2$, $\rho_{(2)} = 0.8$, $\rho_{(3)} = 0.5$, $\eta_{0(21)} = 0.5$, $\eta_{1(21)} = 0.3$, $\eta_{0(31)} = 1$, $\eta_{1(31)} = 0.6$, $\eta_{0(32)} = 1.5$, and $\eta_{1(32)} = 0.9$ to completely specify $\boldsymbol{Q}_w$ for each of the 6 models. For each M_i, we generated 50 datasets by first generating $\boldsymbol{w} \sim N(\mathbf{0}, \boldsymbol{Q}_w)$ and then generating y_{ij}'s from (2.4) using the specifications described above.

2.3.4.2 Comparisons between MDAGAR and GMCAR

In our first experiment, we analyzed the 85 replicated datasets using (2.7) with

$$p(\boldsymbol{\rho}) \times p(\boldsymbol{\eta}) \propto \prod_{i=1}^{q=2} \{Unif(\rho_i \,|\, 0, 1)\} \times N(\boldsymbol{\eta}_{21} \,|\, \mathbf{0}, 0.01\boldsymbol{I}_2) \,, \tag{2.19}$$

where $\boldsymbol{\eta}_{21} = (\eta_{021}, \eta_{121})^\top$ and $Unif$ is the Uniform density. Prior specifications are completed by setting $a_\tau = 2$, $b_\tau = 8$, $a_\sigma = 2$, $b_\sigma = 0.4$, $\boldsymbol{\mu}_\beta = \mathbf{0}$, $\boldsymbol{V}_\beta = 1000\boldsymbol{I}$ in (2.7). Note that the same set of priors were used for both MDAGAR and GMCAR as they have the same number of parameters with similar interpretations.

We compare models using the Widely Applicable Information Criterion (WAIC)[Wat10, GHV14] and a model comparison score D based on a balanced loss function for replicated data [GG98]. Both WAIC and D reward goodness of fit and penalize model complexity. Details on how these metrics are computed are provided in Appendix 2.3.7.2. In addition, we also computed the average mean squared error (AMSE) of the spatial random effects estimated from each of the 85 data sets. We found the mean (standard deviation) of the AMSEs to be 1.69 (0.034) from the 85 low-correlation datasets, 1.47 (0.030) from the 85 medium-correlation datasets, and 2.35 (0.059) from the 85 high-correlation datasets. The corresponding numbers for GMCAR were 1.83 (0.033), 1.59 (0.031), and 2.14 (0.050), respectively. The MDAGAR tends to have smaller AMSE for low and medium correlations, while GMCAR has lower AMSE when the correlations are high, although the differences

are not significant. We also computed the mean values of WAICs and D scores for each simulated data set. Figure 2.6 plots the values of WAICs ((a)–(c)) and D scores ((d)–(f)) for the 85 data sets corresponding to each of the three correlation settings. Here, MDAGAR outperforms GMCAR in all three correlation settings with respect to both WAICs and D scores.

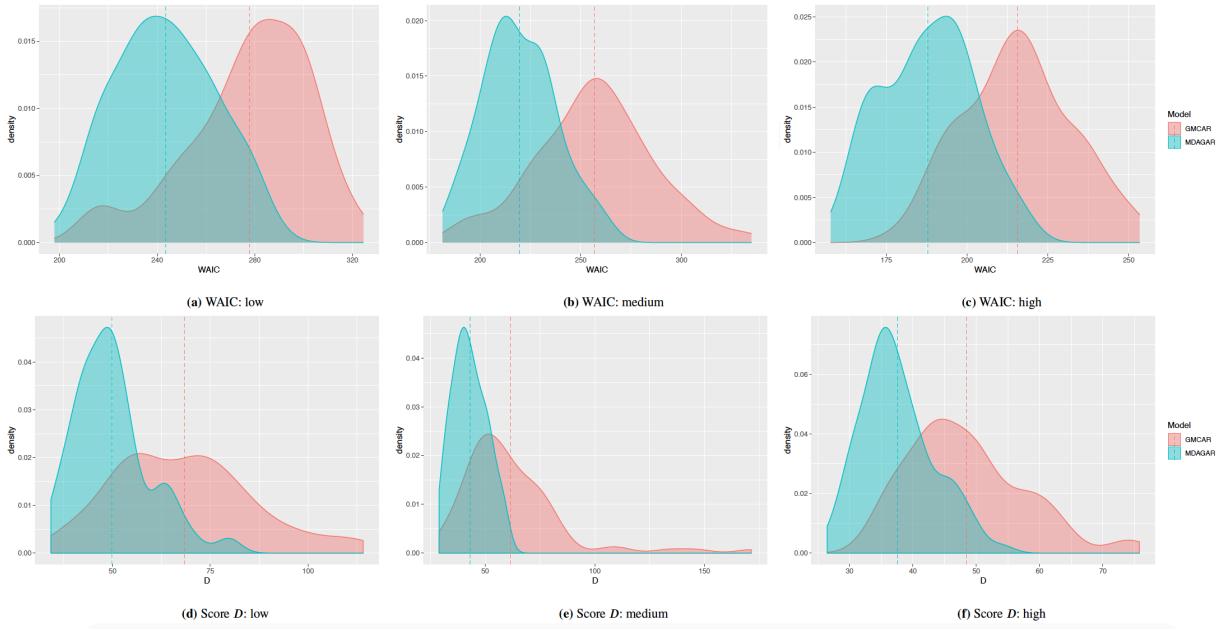


Figure 2.6: Density plots for WAICs and D scores over 85 datasets. Density plots of WAIC for MDAGAR (blue) and GMCAR (red) models with low, medium and high correlation are shown in (a), (b) and (c) respectively, while (d)–(f) are the corresponding density plots for D scores. The dotted vertical line shows the mean for WAIC and D in each plot.

Figure 2.7 presents scatter plots for the true values (x axis) of spatial random effects against their posterior estimates (y axis). To be precise, each panel plots $85 \times 48 \times 2 = 8160$ true values of the elements of the 96×1 vector $\boldsymbol{w}$ for 85 datasets against their corresponding posterior estimates. We see strong agreements between the true values and their estimates for both MDAGAR and GMCAR. The agreement is more pronounced for the datasets corresponding to medium and high correlations. For the low-correlation datasets, MDAGAR still exhibits strong agreement which is better than GMCAR.

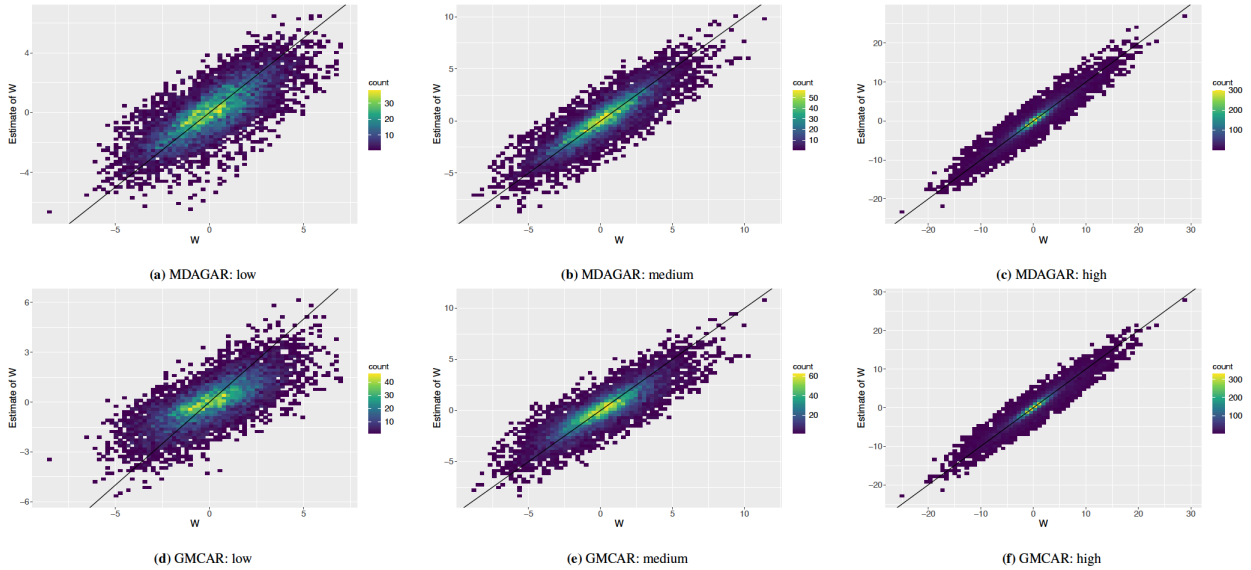


Figure 2.7: Scatter plots for estimates of spatial random effects (y axis) against the true values (x axis) with 45° lines over 85 datasets: (a)–(c) are estimates from MDAGAR model with low, medium and high correlation, while (d)–(f) are the corresponding estimates from GMCAR.

We computed $D_{KL}(N(\mathbf{0}, \boldsymbol{Q}_{true})||N(\mathbf{0}, \boldsymbol{Q}_w)) = \frac{1}{2}\left[\log\left(\frac{\det(\boldsymbol{Q}_{true})}{\det(\boldsymbol{Q}_w)}\right) + \text{tr}(\boldsymbol{Q}_w\boldsymbol{Q}_{true}^{-1}) - qk\right]$, which is the Kullback-Leibler Divergence between the model for $\boldsymbol{w}$ with the true generative precision matrix ($\boldsymbol{Q}_{true}$) and those with MDAGAR and GMCAR precisions ($\boldsymbol{Q}_w$). Using the posterior samples in the precision matrix, we evaluate the posterior probability that $D_{KL}(N(\mathbf{0}, \boldsymbol{Q}_{true})||N(\mathbf{0}, \boldsymbol{Q}_{MDAGAR}))$ is smaller than $D_{KL}(N(\mathbf{0}, \boldsymbol{Q}_{true})||N(\mathbf{0}, \boldsymbol{Q}_w))$. Figure 2.8 depicts a density plot of these probabilities over the 85 data sets. When correlations are low and medium, the MDAGAR has a mean probability of around 69% to be closer to the true model than the GMCAR, while for high correlations GMCAR excels with an average probability of 72% to be closer to the true model. These findings are consistent with the results of AMSE, where the GMCAR tended to perform better when the correlations are high.

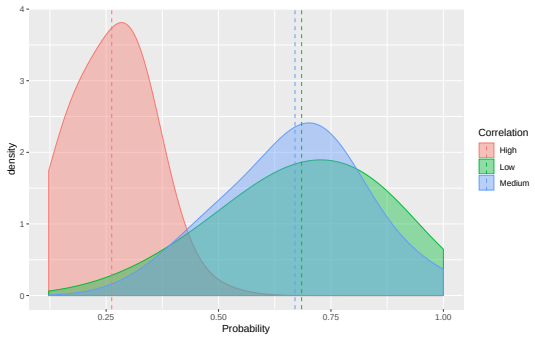


Figure 2.8: Density plots for probability that the KL-divergence between the MDAGAR and the true model is smaller than that between GMCAR and the true model with three levels of correlation for two diseases: low (purple), medium (green) and high (red).

2.3.4.3 Analyses using different orderings for spatial units

The MDAGAR model in Section 2.3.4.2 is analyzed using an ordering of spatial units (counties) from the southwest to the northeast. Here, we repeat the analysis for the MDAGAR model using three other orderings that start in the southeast, northwest and northeast, respectively. We present results from these differently ordered DAGAR models using the 85 low-correlation simulated datasets. For the random effects, the mean (standard deviation) of the AMSEs for three different orderings (southeast, northwest and northeast) are 1.61 (0.029), 1.28 (0.026) and 1.43 (0.027), respectively, without significantly differing from the original ordering in Section 2.3.4.2.

Figure 2.9 plots the densities of mean WAICs, D scores and $D_{KL}(p(\boldsymbol{y}_{true})||p(\boldsymbol{y}))$ over the 85 datasets for the MDAGAR model using three different orderings and the original ordering in Section 2.3.4.2. In computing $D_{KL}(p(\boldsymbol{y}_{true})||p(\boldsymbol{y}))$, we specify $p(\boldsymbol{y}_{true}) = N(\boldsymbol{X}\boldsymbol{\beta}_{true} + \boldsymbol{w}_{true}, diag(\boldsymbol{\sigma}_{true}) \otimes \boldsymbol{I}_n)$, which is the density of the true $\boldsymbol{y}$ and $p(\boldsymbol{y}) = N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}, diag(\boldsymbol{\sigma}) \otimes \boldsymbol{I}_n)$ is the density for $\boldsymbol{y}$ from MDAGAR. We find that the ordering does not have significant impact on model fitting as the density plots for the four orderings almost overlap with each other. These findings are consistent with results, theoretical and empirical, for univariate DAGAR models that the ordering has little impact [DBH19].

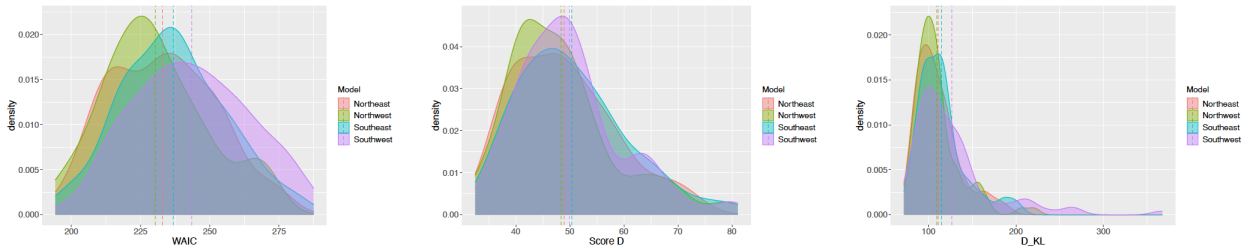


Figure 2.9: Density plots for WAICs, D scores and $D_{KL}(p(\boldsymbol{y}_{true})||p(\boldsymbol{y}))$ over 85 datasets for the MDAGAR model using four different orderings: northeast (red), northwest (green), southeast (blue) and southwest (purple). The dotted vertical line shows the mean for each plot.

2.3.4.4 Model selection for different disease orders

We now evaluate the effectiveness of the method in Section 2.3.3 at selecting the model with the correct ordering of diseases. We used the `bridgesampling` package in `R` to compute $p(M_i \mid \boldsymbol{y}^{(n)}) = \max_{t=1,\ldots,6} p(M_t \mid \boldsymbol{y}^{(n)})$ for each of $n = 50 \times 6$ data sets generated as described in Section 2.3.4.1. Table 2.3 presents the probability of each model being selected for different true model scenarios. The probability of selecting the true model is shown in bold along the diagonal. Our experiment reveals that bridge sampling is extremely effective at choosing the correct order. It is able to identify the correct order between 78% to 90%, which is substantially larger than any of the probability of choosing any of the misspecified models.

Table 2.3: Proportion of times ($\pi(M_i)$) bridge sampling chose the model with the correct order out of the 50 data sets with that order.

True model	$\pi(M_1)$	$\pi(M_2)$	$\pi(M_3)$	$\pi(M_4)$	$\pi(M_5)$	$\pi(M_6)$
M_1	**0.90**	0.00	0.10	0.00	0.00	0.00
M_2	0.00	**0.86**	0.00	0.00	0.14	0.00
M_3	0.14	0.00	**0.86**	0.00	0.00	0.00
M_4	0.00	0.00	0.00	**0.90**	0.00	0.10
M_5	0.00	0.22	0.00	0.00	**0.78**	0.00
M_6	0.00	0.00	0.00	0.16	0.00	**0.84**

2.3.5 Multiple Cancer Analysis from SEER

We now turn to analyzing an areal dataset using the MDAGAR model for all four different cancers: lung, esophagus, larynx and colorectal. The dataset extracted from the SEER*Stat database consists of the four cancers, where the outcome is the 5-year average age-adjusted incidence rates (age-adjusted to the 2000 U.S. Standard Population) per 100,000 population in the years from 2012 to 2016 across 58 counties in California, USA, as mapped in Figure 2.10. The maps exhibit preliminary evidence of correlation across space and among cancers. Cutoffs for the different levels of incidence rates are quantiles for each cancer. For all four cancers, incidence rates are relatively higher in counties concentrated in the middle northern areas including Shasta, Tehama, Glenn, Butte and Yuba than those other areas. In general, northern areas have higher incidence rates than in the south. This is especially pronounced for lung cancer and esophageal cancer. For larynx cancer, while the highest incidence rates are in the northwest (Del Norte and Siskiyou counties), the incidence rates in the south are also at somewhat higher levels. For colorectal cancer, the edge areas at the bottom also exhibit high incidence rates.

As an exploratory tool to assess associations among the cancers, we calculate Pearson's correlation for each pair of cancers by regarding incidence rates in different counties as independent samples and find Pearson's correlation coefficient between the incidence of lung cancer and those of esophageal, larynx and colorectal cancers to be 0.55, 0.46 and 0.46, respectively. Meanwhile, the correlation between esophageal and larynx cancer is 0.27. Next, to explore the spatial association for each disease, we calculate Moran's I based upon rth order neighbors for each cancer and plot the areal correlogram [BCG14]. Defining distance intervals, $(0, d_1], (d_1, d_2], (d_2, d_3], \ldots$, the rth order neighbors refer to units with distance in $(d_{r-1}, d_r]$, i.e. within distance d_r but separated by more than d_{r-1}. The distance is the Euclidean distance from an Albers map projection of California. As shown in Figure 2.11, lung, esophageal and colorectal cancers all present spatial patterns that initially diminish with increasing r and eventually flatten close to 0. Overall, counties with similar levels of

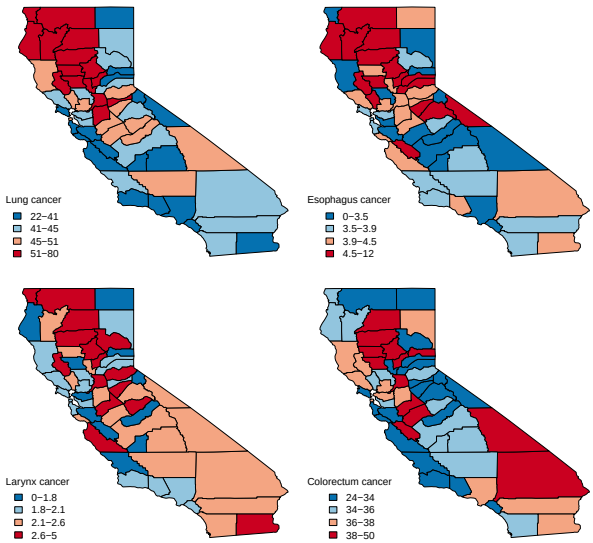


Figure 2.10: Maps of 5-year average age-adjusted incidence rates per 100,000 population for lung, esophagus, larynx and colorectal cancer in California, 2012 – 2016.

incidence rates tend to depict some spatial clustering.

We analyze this data set and separate the spatial correlation for each cancer from association among cancers using (2.7) with the following prior specification

$$p(\boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\tau}, \boldsymbol{\sigma}, \boldsymbol{w}) = \prod_{i=1}^{q} Unif(\rho_i \,|\, 0, 1) \times \prod_{i=2}^{q} \prod_{j=1}^{i-1} N(\boldsymbol{\eta}_{ij} \,|\, 0, 0.01\boldsymbol{I}_2) \times \prod_{i=1}^{q} N(\boldsymbol{\beta}_i \,|\, 0, 0.001\boldsymbol{I})$$
$$\times \prod_{i=1}^{q} IG(1/\tau_i \,|\, 2, 0.1) \times \prod_{i=1}^{q} IG(\sigma_i^2 \,|\, 2, 1) \times N(\boldsymbol{w} \,|\, \boldsymbol{0}, \boldsymbol{Q}_{\boldsymbol{w}}) \,. \tag{2.20}$$

We also discuss a "case 2" excluding the risk factor.

We include the same covariates as Section 2.3.2. Spatial patterns in the map of adult cigarette smoking rates, shown in Figure 2.12, are similar to the incidence of cancers, especially lung and esophageal cancers, the highest smoking rates are concentrated in the north. While some central California counties (e.g., Stanislaus, Tuolumne, Merced, Mariposa, Fresno and Tulare) also exhibit high rates, although there is clearly less spatial clus-

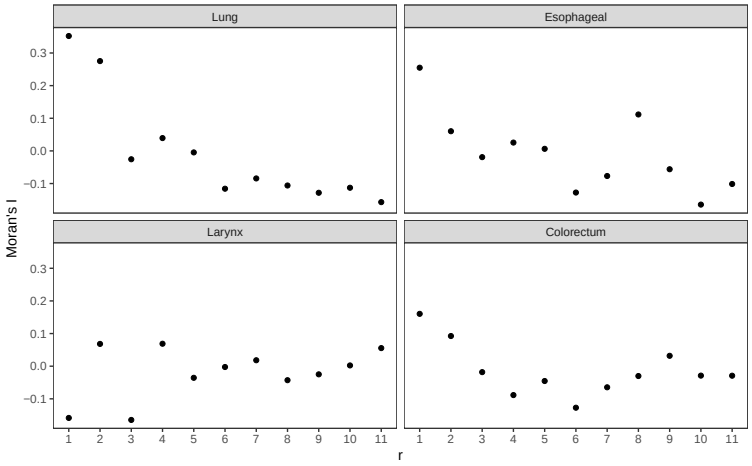


Figure 2.11: Moran's I of rth order neighbors for lung, esophageal, larynx and colorectal cancer.

tering of the high rates than in the north.

Since the order of cancers in the DAG specify the model, we fit all $4! = 24$ models using (2.7) and compute the marginal likelihoods using bridge sampling (Section 2.3.3). By setting the prior model probabilities as $p(M = M_t) = \frac{1}{24}$ for $t = 1, 2, \ldots, 24$, we compute the posterior model probabilities using (2.13). These are presented in Table 2.4. We obtain Bayesian model averaged (BMA) estimates using (2.18) with the weights in Table 2.4. Among all models, model M_{10} is selected as the best model with the largest posterior probability 0.577 and the corresponding conditional structure is [esophageal] $\times$ [larynx | esophageal] $\times$ [colorectal | esophageal, larynx] $\times$ [lung | esophageal, larynx, colorectal]. Table 2.5 is a sum-

Table 2.4: The posterior model probabilities for 24 models.

$p(M_1 \mid \boldsymbol{y})$	$p(M_2 \mid \boldsymbol{y})$	$p(M_3 \mid \boldsymbol{y})$	$p(M_4 \mid \boldsymbol{y})$	$p(M_5 \mid \boldsymbol{y})$	$p(M_6 \mid \boldsymbol{y})$	$p(M_7 \mid \boldsymbol{y})$	$p(M_8 \mid \boldsymbol{y})$
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$p(M_9 \mid \boldsymbol{y})$	$p(M_{10} \mid \boldsymbol{y})$	$p(M_{11} \mid \boldsymbol{y})$	$p(M_{12} \mid \boldsymbol{y})$	$p(M_{13} \mid \boldsymbol{y})$	$p(M_{14} \mid \boldsymbol{y})$	$p(M_{15} \mid \boldsymbol{y})$	$p(M_{16} \mid \boldsymbol{y})$
0.000	**0.577**	0.000	0.000	0.000	0.000	0.342	0.079
$p(M_{17} \mid \boldsymbol{y})$	$p(M_{18} \mid \boldsymbol{y})$	$p(M_{19} \mid \boldsymbol{y})$	$p(M_{20} \mid \boldsymbol{y})$	$p(M_{21} \mid \boldsymbol{y})$	$p(M_{22} \mid \boldsymbol{y})$	$p(M_{23} \mid \boldsymbol{y})$	$p(M_{24} \mid \boldsymbol{y})$
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002

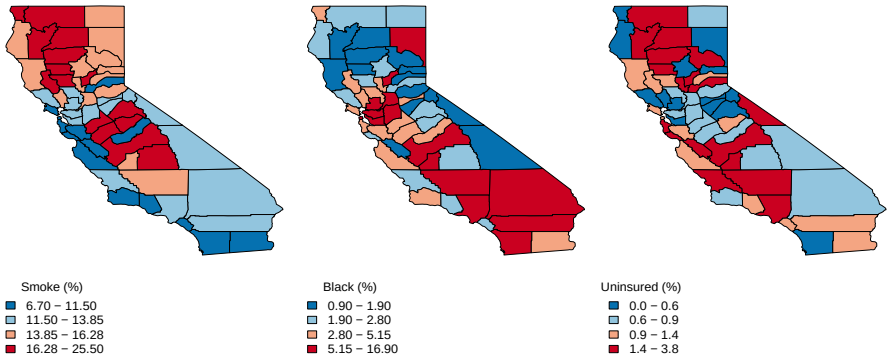


Figure 2.12: Important county-level covariates with significant effects: adult cigarette smoking rates (left), percentage of black residents (middle) and uninsured residents (right).

mary of the parameter estimates including regression coefficients, spatial autocorrelation (ρ_i), spatial precision (τ_i) and noise variance (σ_i^2) for each cancer. From M_{10} and BMA, we find the regression slopes for the percentage of smokers and uninsured residents are significantly positive and negative, respectively, for esophageal cancer. The negative association between percentage of uninsured and esophageal cancer may seem surprising, but is likely a consequence of spatial confounding with counties exhibiting low incidence rates for esophageal cancer having a relatively large number of uninsured residents (see top right in 2.2 and the right most figure in 2.12). Since esophageal cancer has low incidence rates, this association could well be spurious due to spatial confounding. Percentage of smokers is, unsurprisingly, found to be a significant risk factor for lung cancer, while the percentage of blacks seems to be significantly associated with elevated incidence of larynx cancer. In addition, we tend to see that percentage of population below the poverty level has a pronounced association with higher rates of lung and esophageal cancer.

Table 2.5: Posterior means (95% credible intervals) for parameters estimated from M_{10} and BMA estimates for regression coefficients only for the SEER four cancer dataset.

Parameters	Model	Esophageal	Larynx	Colorectal	Lung
Intercept	M_{10}	16.76 (4.06, 29.56)	6.37 (-1.16, 13.89)	19.16 (-11.94, 49.72)	28.68 (-18.3, 74.93)
	BMA	15.87 (2.92, 28.63)	6.85 (-0.71, 14.38)	18.21 (-14.03, 49.07)	28.25 (-18.12, 74.52)
Smokers (%)	M_{10}	**0.25 (0.12, 0.37)**	0.04 (-0.03, 0.12)	0.23 (-0.12, 0.57)	**0.81 (0.08, 1.62)**
	BMA	**0.23 (0.10, 0.36)**	0.05 (-0.03, 0.12)	0.22 (-0.13, 0.58)	**0.80 (0.08, 1.59)**
Young (%)	M_{10}	-0.12 (-0.31, 0.07)	-0.07 (-0.18, 0.04)	0.27 (-0.2, 0.76)	-0.08 (-0.90, 0.74)
	BMA	-0.11 (-0.3, 0.08)	-0.08 (-0.19, 0.03)	0.29 (-0.18, 0.78)	-0.01 (-0.86, 0.82)
Old (%)	M_{10}	-0.11 (-0.25, 0.04)	-0.05 (-0.14, 0.03)	0.10 (-0.28, 0.48)	-0.09 (-0.81, 0.67)
	BMA	-0.10 (-0.25, 0.05)	-0.05 (-0.14, 0.03)	0.10 (-0.29, 0.49)	-0.08 (-0.79, 0.66)
Edu (%)	M_{10}	0.02 (-0.08, 0.12)	-0.02 (-0.08, 0.04)	0.16 (-0.12, 0.43)	-0.20 (-0.75, 0.31)
	BMA	0.02 (-0.09, 0.12)	-0.02 (-0.07, 0.04)	0.15 (-0.14, 0.42)	-0.24 (-0.79, 0.27)
Unemp (%)	M_{10}	-0.13 (-0.29, 0.03)	0.01 (-0.08, 0.10)	-0.09 (-0.54, 0.37)	0.60 (-0.47, 1.55)
	BMA	-0.12 (-0.28, 0.05)	0.01 (-0.08, 0.1)	-0.08 (-0.54, 0.38)	0.61 (-0.43, 1.56)
Black (%)	M_{10}	0.14 (-0.06, 0.34)	**0.14 (0.03, 0.26)**	-0.16 (-0.73, 0.39)	0.15 (-1.06, 1.29)
	BMA	0.13 (-0.07, 0.33)	**0.15 (0.03, 0.27)**	-0.18 (-0.75, 0.39)	0.14 (-1.02, 1.25)
Male (%)	M_{10}	-0.04 (-0.17, 0.09)	0.00 (-0.07, 0.08)	0.24 (-0.12, 0.60)	0.14 (-0.57, 0.79)
	BMA	-0.04 (-0.17, 0.09)	0 (-0.07, 0.08)	0.24 (-0.12, 0.62)	0.14 (-0.55, 0.82)
Uninsured (%)	M_{10}	**-0.24 (-0.44, -0.04)**	-0.08 (-0.20, 0.04)	0.07 (-0.44, 0.58)	0.01 (-0.82, 0.86)
	BMA	**-0.23 (-0.42, -0.02)**	-0.08 (-0.2, 0.04)	0.09 (-0.42, 0.61)	0 (-0.81, 0.82)
Poverty (%)	M_{10}	0.30 (-0.24, 0.84)	0.20 (-0.12, 0.51)	-0.06 (-1.51, 1.45)	0.85 (-2.15, 3.85)
	BMA	0.32 (-0.23, 0.87)	0.2 (-0.12, 0.51)	-0.08 (-1.54, 1.42)	0.8 (-2.14, 3.75)
ρ_{cancer}	M_{10}	0.25 (0.01, 1.00)	0.33 (0.01, 0.96)	0.50 (0.03, 0.97)	0.52 (0.03, 0.99)
τ_{cancer}	M_{10}	25.27 (5.08, 61.57)	27.60 (8.05, 60.42)	19.97 (3.06, 55.61)	20.31 (1.77, 55.92)
σ^2_{cancer}	M_{10}	1.67 (1.11, 2.47)	0.49 (0.28, 0.75)	8.22 (1.09, 14.23)	1.19 (0.18, 5.21)

Recall that ρ_1 is the residual spatial autocorrelation for esophageal cancer after accounting for the explanatory variables, while ρ_i for $i = 2, 3, 4$ are residual spatial autocorrelations after accounting for the explanatory variables and the preceding cancers in the model M_{10}. From Table 2.5 we see that esophageal cancer exhibits relatively weaker spatial autocorrelation, while the residual spatial autocorrelations for larynx and colorectal cancers after accounting for preceding cancers are both at moderate levels of around 0.5. Similarly for the spatial precision τ_i, larynx appears to have the smallest conditional variability while that for colorectal and lung are slightly larger.

For the posterior mean incidence rates and spatial random effects w_{ij}, we present estimates from model M_{10} and BMA. Figure 2.13 (a) and (b) are maps of posterior mean spatial random effects and model fitted incidence rates for four cancers obtained from BMA, while Figure 2.14 (a) and (b) show maps of those from model M_{10}. The posterior mean incidence rates from BMA and M_{10} are in accord with each other, and both present DAGAR-smoothed

versions of the original patterns in Figure 2.2. For posterior means of spatial random effects, in general, the estimates from M_{10} are similar to model averaged estimates, especially for lung and colorectal cancers, exhibiting relatively large positive values in the northern counties, where the incidence rates are high. However, for esophageal and larynx cancers we see slight discrepancies between M_{10} and BMA in the north. The BMA estimates produce larger positive random effects, ranging between $0.1 - 0.5$, in most counties, while M_{10} produces estimates between $0 - 0.1$ for esophageal cancer. More counties with random effects larger than 0.1 are estimated from M_{10} for larynx cancer. We believe this is attributable, at least in part, to another competitive model, $M_{15} = [\text{larynx}] \times [\text{esophagus} \,|\, \text{larynx}] \times [\text{lung} \,|\, \text{larynx}, \text{esophagus}] \times [\text{colorectal} \,|\, \text{larynx}, \text{esophagus}, \text{lung}]$ (posterior probability 0.342), which contributes to the BMA. On the other hand, the effects of some important county-level covariates play an essential role in the discrepancy between the estimates of random effects and model fitted incidence rates for each cancer.

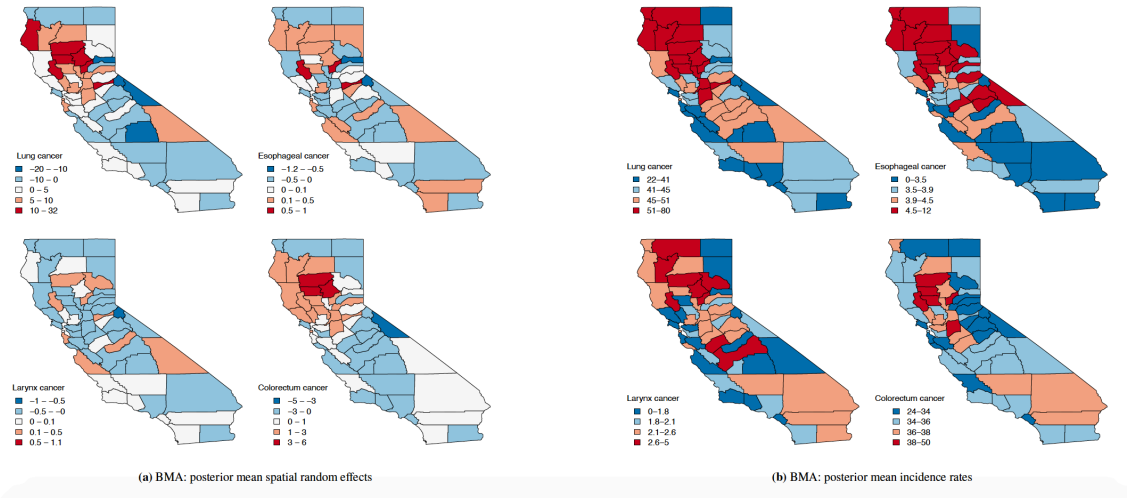
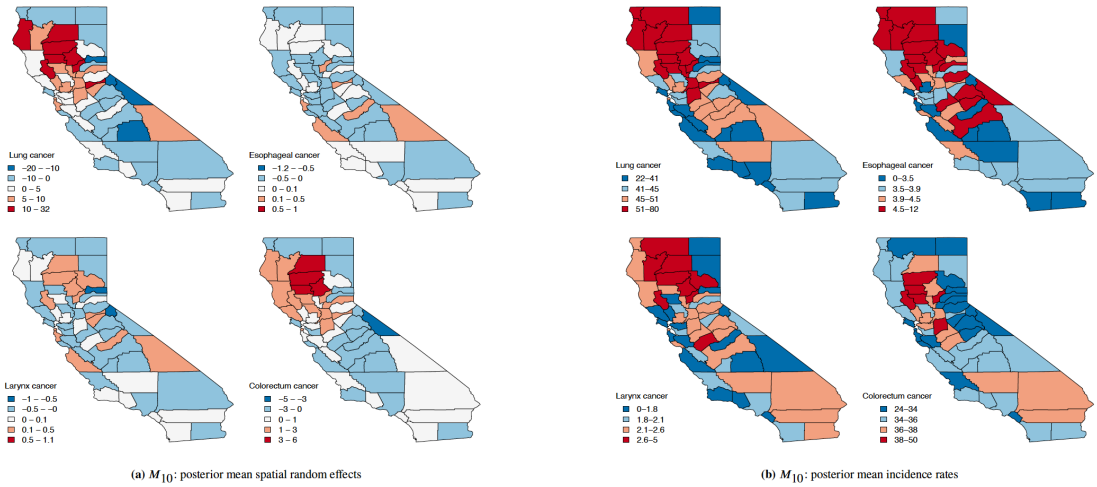


Figure 2.13: Maps of posterior results using BMA for lung, esophagus, larynx and colorectal cancer in California including (a) posterior mean spatial random effects and (b) posterior mean incidence rates.

Recall that $\eta_{0ii'}$ and $\eta_{1ii'}$ reflect the associations among cancers that can be attributed to spatial structure. Specifically, larger values of $\eta_{0ii'}$ will indicate inherent associations



(a) M_{10}: posterior mean spatial random effects

(b) M_{10}: posterior mean incidence rates

Figure 2.14: Maps of posterior results using the highest probability model M_{10} for lung, esophagus, larynx and colorectal cancer in California including (a) posterior mean spatial random effects and (b) posterior mean incidence rates.

unrelated to spatial structure, while the magnitude of $\eta_{1ii'}$ reflects associations due to spatial structure. Figure 2.15 presents posterior distributions of $\boldsymbol{\eta}$ for all pairs of cancers. We see from the distribution of η_{043} that there is a pronounced non-spatial component in the association between lung and colorectal cancers. Similar, albeit somewhat less pronounced, non-spatial associations are seen between larynx and esophageal cancers and between lung and larynx cancers. Analogously, the posterior distributions for η_{143} and η_{132} tend to have substantial positive support suggesting substantial spatial cross-correlations between lung and colorectal cancers and between colorectal and larynx cancers. Interestingly, we find negative support in the posterior distributions for η_{121} and η_{142}. The negative mass implies that the covariance among cancers within a region is suppressed by strong dependence with neighboring regions. This seems to be the case for associations between lung and esophageal cancers and between lung and larynx cancers.

We also present supplementary analysis that excludes adult smoking rates from the covariates, which we refer to as "Case 2". Excluding the risk factor adult cigarette smoking rates, we only include county attributes described in Section 2.3.5 as covariates. Among

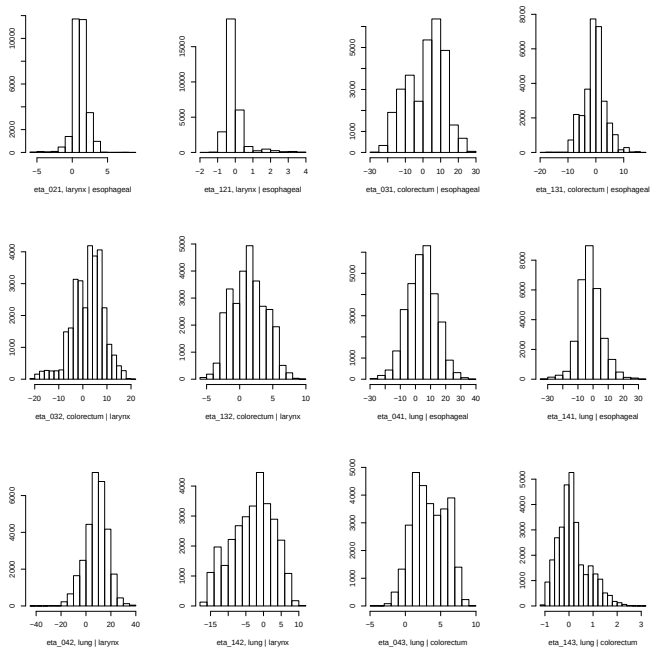


Figure 2.15: Posterior distributions of $\boldsymbol{\eta}$ for all pairs of cancers.

24 models, model M_{16} exhibits dominated best performance with a posterior probability of 0.999 and the corresponding conditional structure is [larynx] $\times$ [esophagus $|$ larynx] $\times$ [colorectal $|$ larynx, esophagus] $\times$ [lung $|$ larynx, esophagus, colorectal]. Table 2.6 is a summary of the parameter estimates for each cancer. From M_{16}, we find that the regression slope for the percentage of blacks and unemployed residents are significantly positive for larynx and lung cancer respectively. The larynx cancer exhibits weaker spatial autocorrelation while the residual spatial autocorrelation for the other three cancers after accounting for preceding cancers are at moderate levels. For spatial precision τ_i, larynx random effects still have the smallest variability while the conditional variability for colorectal and lung cancers are slightly larger.

Figure 2.16 shows estimated correlations between pairwise cancers in each of the 58 counties. The top row presents the correlations including smoking rates ("Case 1") as has been analyzed here. The bottom row presents the corresponding maps for "Case 2". Interestingly, accounting for smoking rates substantially diminishes the associations among

esophageal, colorectal and lung cancers. These are significantly associated in "Case 2" but only lung and colorectal retain their significance after accounting for smoking rates.

Table 2.6: Posterior means (95% credible intervals)for parameter estimated from M_{16} for Case 2 (excluding smoking rates in covariates).

Parameters	Larynx	Esophageal	Colorectal	Lung
Intercept	6.75 (-0.58, 14.00)	11.14 (-1.70, 24.05)	18.89 (-10.37, 48.12)	24.18 (-22.71, 68.75)
Young(%)	-0.09 (-0.20, 0.02)	-0.09 (-0.29, 0.11)	0.27 (-0.19, 0.74)	0.04 (-0.75, 0.86)
Old(%)	-0.04 (-0.12, 0.04)	0.00 (-0.15, 0.16)	0.13 (-0.23, 0.49)	0.15 (-0.49, 0.91)
Edu(%)	-0.02 (-0.08, 0.04)	-0.02 (-0.13, 0.09)	0.12 (-0.13, 0.38)	-0.34 (-0.82, 0.15)
Unemp(%)	0.04 (-0.03, 0.12)	0.06 (-0.08, 0.20)	0.10 (-0.26, 0.45)	**1.21 (0.55, 1.89)**
Black(%)	**0.15 (0.03, 0.27)**	0.10 (-0.12, 0.32)	-0.20 (-0.75, 0.33)	0.06 (-1.03, 1.13)
Male(%)	-0.01 (-0.08, 0.07)	-0.07 (-0.21, 0.06)	0.18 (-0.16, 0.52)	0.01 (-0.59, 0.60)
Uninsured(%)	-0.07 (-0.19, 0.04)	-0.13 (-0.33, 0.07)	0.10 (-0.37, 0.58)	0.11 (-0.70, 0.95)
Poverty(%)	0.21 (-0.11, 0.53)	0.40 (-0.20, 1.02)	0.03 (-1.38, 1.45)	0.84 (-2.14, 3.52)
ρ_i	0.25 (0.01, 0.91)	0.49 (0.02, 0.97)	0.43 (0.02, 0.94)	0.50 (0.03, 0.98)
τ_i	44.04 (15.89, 90.23)	24.55 (5.06, 61.33)	18.25 (1.39, 51.15)	19.68 (2.00, 55.07)
σ_i^2	0.56 (0.37, 0.84)	1.52 (0.88, 2.36)	9.85 (6.48, 14.63)	0.93 (0.18, 3.63)

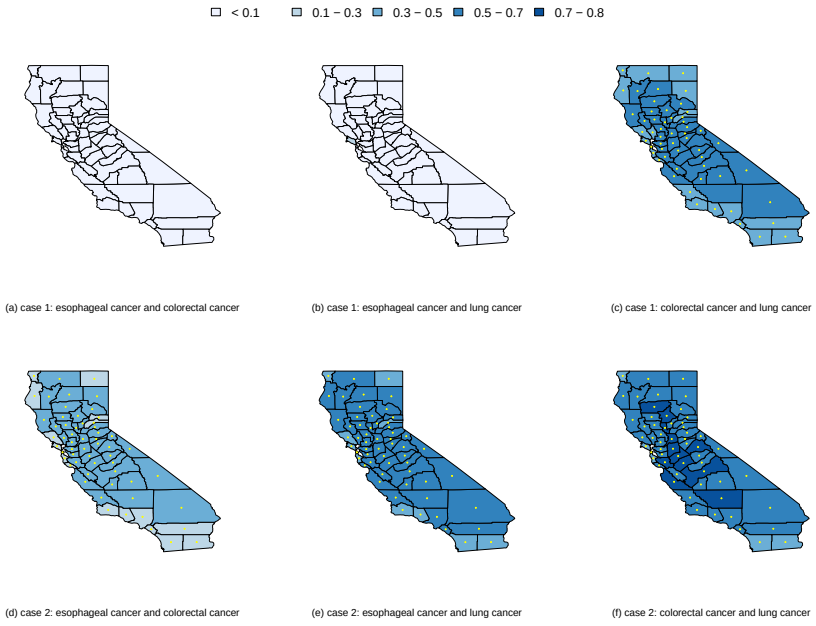


Figure 2.16: Estimated correlation between the incidence of pairwise cancers in each of 58 counties of California for Case 1 vs. Case 2: (a) case 1: esophageal and colorectal cancer, (b) case 1: esophageal and lung cancer, (c) case 1: colorectal and lung cancer, (d) case 2: esophageal and colorectal cancer, (e) case 2: esophageal and lung cancer, (f) case 2: colorectal and lung cancer. Maps (a)-(c) exhibit estimated correlations for Case 1, and (d) - (f) are for Case 2. Yellow points indicate significant correlations. Note: Maps for larynx cancer are not shown due to non-significant correlation with any of the other three cancers.

2.3.6 Summary

This chapter developed a conditional multivariate "MDAGAR" model to estimate spatial correlations for multiple correlated diseases based on a currently proposed class of DAGAR models for univariate disease mapping, as well as providing better interpretation of the association among diseases. An application to bivariate case using the BDAGAR model analyzing incidence rates from lung and esophagus cancer retains the interpretation of DAGAR models clearly separating the spatial correlation for each cancer from any inherent or endemic association between the two cancers. The BDAGAR model can still be efficiently computed using MCMC algorithms. The analysis demonstrates the efficiency of BDAGAR and its improved performance, as measured by WAIC, over existing alternatives such as the GMCAR models. In fact, it has been reported that DAGAR tended to outperform CAR in univariate models [DBH19]. It is, therefore, not unexpected that BDAGAR will outperform the bivariate CAR models.

Then the example of BDAGAR is generalized to MDAGAR for multivariate disease mapping. We demonstrate that MDAGAR tends to outperform GMCAR when association between spatial random effects for different diseases is weak or moderate. Inference is competitive when associations are strong. MDAGAR retains the interpretability of spatial autocorrelations, as in univariate DAGAR, separating the spatial correlation for each disease from any inherent or endemic association among diseases. While MDAGAR, like all DAG based models, is specified according to a fixed order of the diseases, we show that a bridge sampling algorithm can effectively choose among the different orders and provide Bayesian model averaged inference in a computationally efficient manner. The data analysis for four cancers reveals that correlations between incidence rates for different cancers are impacted by covariates. For example, eliminating adult cigarette smoking rates produces similar spatial patterns for the incidence rates of esophageal, lung and colorectal cancer. In addition, the significant correlation between lung and esophageal cancer, even after accounting for smoking rates, implies other inherent or endemic association such as latent risk factors and

metabolic mechanisms. We also see that the MDAGAR based posterior estimates of the latent spatial effects in Figure 2.13 (a) and 2.14(a) resemble those from an MDAGAR without covariates (Figure 2.17), while the maps for the estimated incidence rates in Figure 2.13 (b) and 2.14 (b) account for the spatial variability of the covariates.

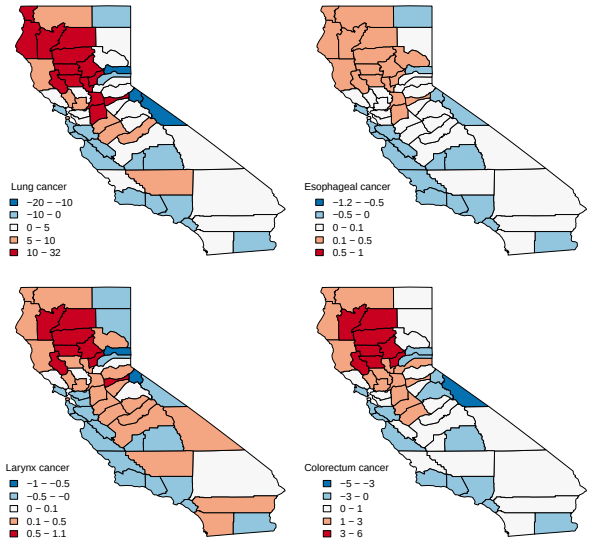


Figure 2.17: Maps of posterior mean spatial random effects (with no covariates) using the same order as M_{10}.

Future challenges will include scalability with very large number of diseases because, as we have seen, the number of models to be fitted grows exponentially with the number of diseases. One way to obviate this issue is to adopt a joint modeling approach analogous to order-free MCAR models [JBC07] that build rich spatial structures from linear transformations of simpler latent variables. For instance, we can develop alternate MDAGAR models by specifying $\boldsymbol{w} = \boldsymbol{\Lambda}\boldsymbol{f}$, where $\boldsymbol{\Lambda}$ is a suitably specified matrix and $\boldsymbol{f}$ is a latent vector whose components follow independent univariate DAGAR distributions as discussed in Section 2.4. This will avoid the order dependence, but the issue of identifying and specifying $\boldsymbol{\Lambda}$ will need to be considered as will the interpretation of disease specific spatial autocorrelations.

2.3.7 Appendix

2.3.7.1 Model Implementation

We outline model implementation for (2.7) using Markov Chain Monte Carlo (MCMC). We update $\{\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_q\}$ using Gibbs steps, while the elements of $\boldsymbol{\rho}$ are updated from their full conditional distributions using Metropolis random walk steps [RC13]. A particularly appealing feature of our proposed MDAGAR model is that the spatial weight parameters $\boldsymbol{\eta} = \{\boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_q\}$ render Gaussian full conditional distributions in addition to the customary Gaussian full conditional distributions for $\boldsymbol{\beta}$ and $\boldsymbol{w}$. As a matter of notational convenience for the derivations that follow, we use $N(\boldsymbol{\mu}, \boldsymbol{V})$ to denote the normal distribution with variance-covariance matrix $\boldsymbol{V}$. This difference from our notation in the main manuscript where we use the precision matrix in the argument of normal distribution.

Full Conditional Distributions The full conditional distribution for each $\boldsymbol{\beta}_i$ is

$$\boldsymbol{\beta}_i | \boldsymbol{y}_i, \boldsymbol{w}_i, \sigma_i^2 \sim N(\boldsymbol{M}_i \boldsymbol{m}_i, \boldsymbol{M}_i) \tag{2.21}$$

where $\boldsymbol{M}_i = \left(\frac{1}{\sigma_i^2}\boldsymbol{X}_i^\top \boldsymbol{X}_i + \frac{1}{\sigma_\beta^2}\boldsymbol{I}_{p_i}\right)^{-1}$ and $\boldsymbol{m}_i = \frac{1}{\sigma_i^2}\boldsymbol{X}_i^\top (\boldsymbol{y}_i - \boldsymbol{w}_i)$. Similarly, the full conditional distribution of each σ_i^2 follows an inverse gamma distribution,

$$\sigma_i^2 | \boldsymbol{y}_i, \boldsymbol{\beta}_i, \boldsymbol{w}_i \sim IG\left(a_\sigma + \frac{k}{2}, b_\sigma + \frac{1}{2}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i - \boldsymbol{w}_i\right)^\top \left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i - \boldsymbol{w}_i\right)\right). \tag{2.22}$$

The full conditional distribution for $\boldsymbol{w}_i$ for each $i = 2, \ldots, q-1$ is

$$
\begin{aligned}
&p\left(\boldsymbol{w}_i \,|\, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{i-1}, \boldsymbol{w}_{i+1}, \boldsymbol{w}_{i+1}, \ldots, \boldsymbol{w}_q, \boldsymbol{y}_i, \boldsymbol{\beta}_i, \sigma_i^2, \boldsymbol{\eta}_i, \ldots, \boldsymbol{\eta}_q, \rho_i, \ldots, \rho_q, \tau_i, \ldots, \tau_q\right) \\
&\qquad \propto \prod_{n=i}^{q} \exp\left\{ -\frac{\tau_n}{2} \left(\boldsymbol{w}_n - \sum_{i'=1}^{n-1} \boldsymbol{A}_{ni'} \boldsymbol{w}_{i'} \right)^\top \boldsymbol{Q}\left(\rho_n\right) \left(\boldsymbol{w}_n - \sum_{i'=1}^{n-1} \boldsymbol{A}_{ni'} \boldsymbol{w}_{i'} \right) \right\} \\
&\qquad\qquad \times \exp\left\{ -\frac{1}{2\sigma_i^2} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i - \boldsymbol{w}_i \right)^\top \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i - \boldsymbol{w}_i \right) \right\} \qquad (2.23)
\end{aligned}
$$

which is equal to $N(\boldsymbol{w}_i \,|\, \boldsymbol{G}_i \boldsymbol{g}_i, \boldsymbol{G}_i)$, where

$$
\boldsymbol{G}_i = \left[\tau_i \boldsymbol{Q}(\rho_i) + \sum_{n=i+1}^{q} \tau_n \boldsymbol{A}_{ni}^\top \boldsymbol{Q}(\rho_n) \boldsymbol{A}_{ni} + \frac{1}{\sigma_i^2} \boldsymbol{I}_k \right]^{-1}
$$

and $\boldsymbol{g}_i = \tau_i \boldsymbol{Q}(\rho_i) \sum_{n=1}^{i-1} \boldsymbol{A}_{in} \boldsymbol{w}_n + \sum_{n=i+1}^{q} \tau_n \boldsymbol{A}_{ni}^\top \boldsymbol{Q}(\rho_n) \left(\boldsymbol{w}_n - \sum_{i'=1, i' \neq i}^{n-1} \boldsymbol{A}_{ni'} \boldsymbol{w}_{i'} \right) + \frac{1}{\sigma_i^2} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i)$.
For $i = 1$ and q, we have

$$
\begin{aligned}
\boldsymbol{w}_1 | \boldsymbol{w}_2, \ldots, \boldsymbol{w}_q, \boldsymbol{y}_1, \boldsymbol{\beta}_1, \sigma_1^2, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\tau} &\sim N(\boldsymbol{G}_1 \boldsymbol{g}_1, \boldsymbol{G}_1) \\
\boldsymbol{w}_q | \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{q-1}, \boldsymbol{y}_q, \boldsymbol{\beta}_q, \sigma_1^2, \boldsymbol{\eta}_q, \rho_q, \tau_q &\sim N(\boldsymbol{G}_q \boldsymbol{g}_q, \boldsymbol{G}_q)
\end{aligned}
$$

, where

$$
\begin{aligned}
\boldsymbol{G}_1 &= \left(\tau_1 \boldsymbol{Q}(\rho_1) + \sum_{n=2}^{q} \tau_n \boldsymbol{A}_{n1}^\top \boldsymbol{Q}(\rho_n) \boldsymbol{A}_{n1} + \frac{1}{\sigma_1^2} \boldsymbol{I}_k \right)^{-1}, \\
\boldsymbol{g}_1 &= \tau_2 \boldsymbol{A}_{21}^\top \boldsymbol{Q}(\rho_2) \boldsymbol{w}_2 + \sum_{n=3}^{q} \tau_n \boldsymbol{A}_{n1}^\top \boldsymbol{Q}(\rho_n) \left(\boldsymbol{w}_n - \sum_{i'=2}^{n-1} \boldsymbol{A}_{ni'} \boldsymbol{w}_{i'} \right) + \frac{1}{\sigma_1^2} (\boldsymbol{y}_1 - \boldsymbol{X}_1 \boldsymbol{\beta}_1), \\
\boldsymbol{G}_q &= \left(\tau_q \boldsymbol{Q}(\rho_q) + \frac{1}{\sigma_q^2} \boldsymbol{I}_k \right)^{-1} \\
\boldsymbol{g}_q &= \tau_q \boldsymbol{Q}(\rho_q) \sum_{n=1}^{q-1} \boldsymbol{A}_{qn} \boldsymbol{w}_n + \frac{1}{\sigma_q^2} (\boldsymbol{y}_q - \boldsymbol{X}_q \boldsymbol{\beta}_q).
\end{aligned}
$$

The full conditional distribution of each τ_i is

$$\tau_1|\boldsymbol{w}_1,\rho_1 \sim G\left(a_{\tau_1}+\frac{k}{2}, b_{\tau_1}+\frac{1}{2}\boldsymbol{w}_1^T\boldsymbol{Q}(\rho_1)\boldsymbol{w}_1\right),$$

$$\tau_i|\boldsymbol{w}_1,\ldots,\boldsymbol{w}_i,\boldsymbol{\eta}_i,\rho_i \sim G\left(a_{\tau_i}+\frac{k}{2}, b_{\tau_i}+\frac{1}{2}\left(\boldsymbol{w}_i-\sum_{i'=1}^{i-1}\boldsymbol{A}_{i,i'}\boldsymbol{w}_{i'}\right)^\top \boldsymbol{Q}(\rho_i)\left(\boldsymbol{w}_i-\sum_{i'=1}^{i-1}\boldsymbol{A}_{i,i'}\boldsymbol{w}_{i'}\right)\right), \quad i=2,3,\ldots,q$$

We now derive the full conditional distributions for the $\boldsymbol{\eta}_i$s. From (2.5) with $i=2$, each element in $\boldsymbol{w}_2$ can be written as $w_{2j}=\eta_{021}w_{1j}+\eta_{121}\sum_{j'\sim j}w_{1j'}+\epsilon_{2j}$, where ϵ_{2j} is the jth element in $\boldsymbol{\epsilon}_2$. To extract $\boldsymbol{\eta}_{21}=(\eta_{021},\eta_{121})^\top$ from the matrix $\boldsymbol{A}_{21}$, $\boldsymbol{A}_{21}\boldsymbol{w}_1$ is rewritten as $\boldsymbol{Z}_1\boldsymbol{\eta}_{21}$ where $\boldsymbol{Z}_1=(\boldsymbol{w}_1,\boldsymbol{\zeta}_1)$ and $\boldsymbol{\zeta}_1=\left(\sum_{j'\sim 1}w_{1j'},\ldots,\sum_{j'\sim k}w_{1j'}\right)^\top$. In general, $\boldsymbol{A}_{ii'}\boldsymbol{w}_{i'}=\boldsymbol{Z}_{i'}\boldsymbol{\eta}_{ii'}$ with $\boldsymbol{Z}_{i'}=(\boldsymbol{w}_{i'},\boldsymbol{\zeta}_{i'})$, where $\boldsymbol{\zeta}_{i'}=\left(\sum_{j'\sim 1}w_{i'j'},\ldots,\sum_{j'\sim k}w_{i'j'}\right)^\top$. Consequently, (5) can be written as $\boldsymbol{w}_i=\boldsymbol{\delta}_i\boldsymbol{\eta}_i+\boldsymbol{\epsilon}_i$, where block matrix $\boldsymbol{\delta}_i=(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_{i-1})$. If $\boldsymbol{\eta}_i\sim N(\boldsymbol{\mu}_i,\boldsymbol{V}_i)$, then the full conditional distribution of $\boldsymbol{\eta}_i$ is

$$p(\boldsymbol{\eta}_i\,|\,\boldsymbol{w}_1,\ldots,\boldsymbol{w}_i,\rho_i)\propto\exp\left\{-\frac{\tau_i}{2}(\boldsymbol{w}_i-\boldsymbol{\delta}_i\boldsymbol{\eta}_i)^\top\boldsymbol{Q}(\rho_i)(\boldsymbol{w}_i-\boldsymbol{\delta}_i\boldsymbol{\eta}_i)\right\} \times\exp\left\{-\frac{1}{2}(\boldsymbol{\eta}_i-\boldsymbol{\mu}_i)^\top\boldsymbol{V}_i^{-1}(\boldsymbol{\eta}_i-\boldsymbol{\mu}_i)\right\}. \tag{2.24}$$

The above is equal to $N(\boldsymbol{\eta}_i\,|\,\boldsymbol{H}_i\boldsymbol{h}_i,\boldsymbol{H}_i)$, where $\boldsymbol{H}_i=\left(\tau_i\boldsymbol{\delta}_i^\top\boldsymbol{Q}(\rho_i)\boldsymbol{\delta}_i+\boldsymbol{V}_i^{-1}\right)^{-1}$ and $\boldsymbol{h}_i=\tau_i\boldsymbol{\delta}_i^\top\boldsymbol{Q}(\rho_i)\boldsymbol{w}_i+\boldsymbol{V}_i^{-1}\boldsymbol{\mu}_i$. For our analysis we set $\boldsymbol{\mu}_i=\boldsymbol{0}$ and $\boldsymbol{V}_i=1000\boldsymbol{I}$.

Metropolis within Gibbs Let $\gamma_i=\log(\frac{\rho_i}{1-\rho_i})$, $\gamma_i\in\mathbb{R}$ and $\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_q)^\top$. The full conditional distribution of $\boldsymbol{\gamma}$ is

$$p(\boldsymbol{\gamma}|\boldsymbol{w},\boldsymbol{\eta}_2,\ldots,\boldsymbol{\eta}_q,\boldsymbol{\tau})\propto p(\boldsymbol{w}|\boldsymbol{\tau},\boldsymbol{\eta}_2,\ldots,\boldsymbol{\eta}_q,\boldsymbol{\rho})\times p(\boldsymbol{\rho})|J|, \tag{2.25}$$

where $p(\boldsymbol{w}|\boldsymbol{\tau}, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_q, \boldsymbol{\rho}) = N(\boldsymbol{w} \,|\, \boldsymbol{G}\boldsymbol{g}, \boldsymbol{G})$, $\boldsymbol{G} = \left(\boldsymbol{Q}_w + \boldsymbol{\Sigma}^{-1}\right)^{-1}$, $\boldsymbol{g} = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)$, $\boldsymbol{\Sigma} = diag(\boldsymbol{\sigma}) \otimes \boldsymbol{I}_k$ and $J = \prod_{i=1}^{q} \rho_i(1-\rho_i)$. Using the formula of transformation, $p(\boldsymbol{\rho})|J|$ is the prior for $\boldsymbol{\gamma}$ and in the right-hand side, $\boldsymbol{\rho}$ can be substituted by $\boldsymbol{\gamma}$ given $\rho_i = \frac{e^{\gamma_i}}{1+e^{\gamma_i}}$.

In our analysis, for each model we ran two MCMC chains for 30,000 iterations each. Posterior inference was based upon 15,000 samples retained after adequate convergence was diagnosed. The MDAGAR model in the simulation examples was programmed in the S language as implemented in the `R` statistical computing environment. All other models were implemented using the `rjags` package available from CRAN `https://cran.r-project.org/web/packages/rjags/`.

2.3.7.2 Simulation

WAIC, AMSE and D score For the simulation studies in Section 2.3.4.2, let $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{w}\}$. The likelihood of each data point $p(y_{ij} \mid \boldsymbol{\theta}) = p\left(y_{ij} \mid \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}_i + w_{ij}, 1/\sigma_i^2\right)$ is needed for calculating WAIC which is defined as

$$WAIC = -2\left(\widehat{lpd} - \hat{p}_{WAIC}\right) \ ,$$

where $\widehat{lpd}$ is computed using posterior samples as the sum of log average predictive density i.e. $\sum_{i=1}^{q}\sum_{j=1}^{k} \log\left(\frac{1}{L}\sum_{\ell=1}^{L} p\left(y_{ij} \mid \boldsymbol{\theta}^{(\ell)}\right)\right)$, $\boldsymbol{\theta}^{(\ell)}$ for $\ell = 1, \ldots, L$ being L posterior samples of $\boldsymbol{\theta}$, and $\hat{p}_{WAIC}$ is the estimated effective number of parameters computed as

$$\sum_{i=1}^{q}\sum_{j=1}^{k} V_{\ell=1}^{L}\left(\log p\left(y_{ij} \mid \boldsymbol{\theta}^{(\ell)}\right)\right)$$

with $V_{\ell=1}^{L}\left(\log p\left(y_{ij} \mid \boldsymbol{\theta}^{(\ell)}\right)\right) = \frac{1}{L-1}\sum_{\ell=1}^{L}\left[\log p\left(y_{ij} \mid \boldsymbol{\theta}^{(\ell)}\right) - \frac{1}{L}\sum_{\ell=1}^{L}\log p\left(y_{ij} \mid \boldsymbol{\theta}^{(\ell)}\right)\right]^2$.

Turning to the D score, we draw replicates y_{ij}, $y_{\text{rep},ij}^{(\ell)} \sim N(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}_i^{(\ell)} + w_{ij}^{(\ell)}, 1/\sigma_i^{2(\ell)})$ and compute $D = G + P$. Here $G = \sum_{i=1}^{q} ||\boldsymbol{y}_i - \bar{\boldsymbol{y}}_{\text{rep},i}||^2$ is a goodness-of-fit measure, where

$\bar{\boldsymbol{y}}_{\text{rep},i}$ is the mean vector with elements $\bar{y}_{\text{rep},ij} = \frac{1}{L}\sum_{\ell=1}^{L} y_{\text{rep},ij}^{(\ell)}$ and $P = \sum_{i=1}^{q}\sum_{j=1}^{k}\sigma_{\text{rep},ij}^2$ is a summary of variance, where $\sigma_{\text{rep},ij}^2$ is the variance of $y_{\text{rep},ij}^{(\ell)}$ for $\ell = 1, \ldots, L$.

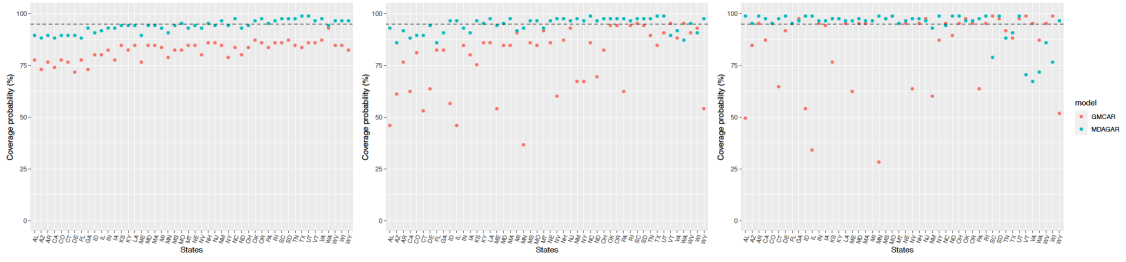
For AMSE, we use w_{ij} as the true value of each random effect and $\hat{w}_{ij}^{(n)}$ is the posterior mean of w_{ij} for the data set n. The estimated AMSE is calculated as $\widehat{AMSE} = \frac{1}{Nqk}\sum_{n=1}^{N}\sum_{i=1}^{q}\sum_{j=1}^{k}\left(\hat{w}_{ij}^{(n)} - w_{ij}\right)^2$ with associated Monte Carlo standard error estimate

$$\widehat{SE}(\widehat{AMSE}) = \sqrt{\frac{1}{(Nqk)(Nqk-1)}\sum_{n=1}^{N}\sum_{i=1}^{q}\sum_{j=1}^{k}\left[\left(\hat{w}_{ij}^{(n)} - w_{ij}\right)^2 - \widehat{AMSE}\right]^2}.$$

Coverage Probability For the simulation studies in Section 2.3.4.2, Figure 2.18 plots coverage probabilities of correlation between two diseases in the same region, given by $corr(w_{1j}, w_{2j}) = cov(w_{1j}, w_{2j})/(\sqrt{var(w_{1j})}\sqrt{var(w_{2j})})$, for MDAGAR and GMCAR. Let $\boldsymbol{Q}(\rho_i)^{-1} = \{d_{ijj'}\}$, we obtain $cov(w_{1j}, w_{2j}) = \tau_1^{-1}(\eta_{021}d_{1jj} + \eta_{121}\sum_{j'\sim j} d_{1jj'})$, $var(w_{1j}) = \tau_1^{-1}d_{1jj}$ and

$$var(w_{2j}) = \tau_1^{-1}[\eta_{021}(\eta_{021}d_{1jj} + \eta_{121}\sum_{j'\sim j} d_{1jj'}) + \eta_{121}\sum_{j'\sim j}(\eta_{021}d_{1jj'} + \eta_{121}\sum_{j''\sim j} d_{1j''j'})] + \tau_2^{-1}d_{2jj}.$$

The MDAGAR performs better in estimating disease correlations in the same region for all scenarios, especially for low and medium correlations with CPs at around 95% in all states.



(a) Low (b) Medium (c) High

Figure 2.18: Coverage probability (%) of $corr(w_{1j}, w_{2j})$, i.e. correlation between two diseases in each state, for MDAGAR (blue) and GMCAR (red).

2.4 A Joint Multivariate DAGAR model for multiple disease mapping

Following the hierarchical construction of the order-free MCAR [JBC07], a joint multivariate DAGAR model is developed to construct the spatial random effects $\boldsymbol{w}_i$ by a linear combination of latent factors $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_i$ for $i \geq 2$ as (1.4), where each $\boldsymbol{f}_i \sim N(0, \boldsymbol{Q}(\rho_i))$ is independently modeled as a univariate DAGAR and $\boldsymbol{Q}(\rho_i)$ are univariate DAGAR precision matrices with $\boldsymbol{B}$ and $\boldsymbol{F}$ as in (2.3) with ρ_i. The joint distribution for $\boldsymbol{w}$ is constructed from $\boldsymbol{w}_1 = a_{11}\boldsymbol{f}_1$ and $\boldsymbol{w}_i = a_{i1}\boldsymbol{f}_1 + a_{i2}\boldsymbol{f}_2 + \cdots + a_{ii}\boldsymbol{f}_i$ for each $i = 2, \ldots, q$, where $a_{ih}, h = 1, \ldots, i$, are coefficients that associate spatial components for different diseases. If $\boldsymbol{F} = (\boldsymbol{f}_1^\top, \ldots, \boldsymbol{f}_i^\top)^\top$ and $\boldsymbol{A}$ be a lower triangular with elements a_{ih}, the joint distribution of $\boldsymbol{w}$ is obtained in matrix form using a $n \times n$ diagonal matrix $\boldsymbol{I}_n$,

$$\boldsymbol{w} = \begin{pmatrix} a_{11}\boldsymbol{I} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ a_{21}\boldsymbol{I} & a_{22}\boldsymbol{I} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q1}\boldsymbol{I} & a_{q2}\boldsymbol{I} & \ldots & a_{qq}\boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \\ \vdots \\ \boldsymbol{f}_q \end{pmatrix} = (\boldsymbol{A} \otimes \boldsymbol{I}_n)\boldsymbol{F}$$

and the covariance matrix of $\boldsymbol{w}$ is

$$\begin{aligned} \boldsymbol{\Sigma}_w &= (\boldsymbol{A} \otimes \boldsymbol{I}_k)\, Cov(\boldsymbol{F})(\boldsymbol{A} \otimes \boldsymbol{I}_k)^\top \\ &= (\boldsymbol{A} \otimes \boldsymbol{I}_k) \left[\bigoplus_{i=1}^{q} \boldsymbol{Q}^{-1}(\rho_i) \right] \left(\boldsymbol{A}^\top \otimes \boldsymbol{I}_k \right) \end{aligned} \tag{2.26}$$

With a shared $\rho_i = \rho$ for all diseases, $\boldsymbol{f}_i \overset{\text{iid}}{\sim} N(\boldsymbol{0}, \boldsymbol{Q}(\rho))$, we obtain a separable covariance matrix $\boldsymbol{\Sigma}_w = (\boldsymbol{A}\boldsymbol{A}^\top) \otimes \boldsymbol{Q}^{-1}(\rho)$ as the Kronecker product of $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\top$ which corresponds to disease dependence and $\boldsymbol{Q}^{-1}(\rho)$ corresponding to spatial association. Henceforth, when $\boldsymbol{\Sigma}_w$ is defined as in (2.26), we denote the distribution of $\boldsymbol{w}$, i.e. $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_w^{-1})$, by MDAGAR$(\rho_1, \ldots, \rho_q, \boldsymbol{\Sigma})$. This MDAGAR model is used to construct the spatial compo-

nents in nonparametric hierarchical models for multivariate difference boundary detection in Section 3. If $\boldsymbol{Q}(\rho_i)$ is specified using a proper CAR structure, the joint distribution of $\boldsymbol{w}$ is presented as MCAR$(\rho_1, \ldots, \rho_q, \boldsymbol{\Sigma})$, i.e. the order-free multivariate CAR model [JBC07].

CHAPTER 3

Multivariate difference boundary detection using nonparametric hierarchical models

3.1 Introduction

For multiple diseases, as shown in Section 2.3, part of the residual from the explanatory variables is captured by the spatial random effect w_{ij} in (2.4). For boundary detection, we define difference boundaries by considering posterior probabilities such as $P(w_{ij} = w_{ij'}|\text{Data}, j \sim j')$ for each disease i and $P(w_{ij} = w_{i'j'}|\text{Data}, j \sim j', i \neq i')$ across different diseases. If the w_{ij}'s are continuous, the probabilities will always be 0 which do not work for boundary detection. To endow the spatial effects with a discrete distribution, two Bayesian nonparametric models using areally-referenced spatial stick-breaking priors (ARSB) and areally-referenced Dirichlet processes (ARDP) [LBH15] are proposed based on Dirichlet process priors for a single disease. Both methods accommodate spatial dependence through DP and model spatial random effects as discrete variables. When there is more than one disease of interest, we also introduce associations among diseases through the nonparametric framework. Considering the computation efficiency and competitive performance, we focus on developing the multivariate model based on ARDP which estimates fewer parameters compared to ARSB.

The ARDP model maintains the marginal distribution of each spatial random effect to be a regular univariate Dirichlet process (DP), denoted as $G^{(j)}(w_j)$, incorporating the spatial dependence between these DPs. The spatial components $\gamma_1, \ldots, \gamma_n$ are jointly distributed as a CAR model and $F^{(1)}(\cdot), \ldots, F^{(n)}(\cdot)$ denote the cumulative distribution functions of

the marginal distributions of γ_j. Marginally, each $F^{(j)}(\gamma_j)$ is uniform $(0,1)$ but they are dependent through $\gamma_1, \ldots, \gamma_k$ introducing spatial dependence between DPs defined below,

$$\begin{aligned}
&\boldsymbol{w} = \{w_j\}_{j=1}^n \sim G_n; G_n = \sum_{u_1,\ldots,u_n} \pi_{u_1,\ldots,u_n} \delta_{\theta_{u_1}} \ldots \delta_{\theta_{u_n}};\\
&\pi_{u_1,\ldots,u_n} = P\left(\sum_{k=1}^{u_1-1} p_k < F^{(1)}(\gamma_1) < \sum_{k=1}^{u_1} p_k, \ldots, \sum_{k=1}^{u_n-1} p_k < F^{(n)}(\gamma_n) < \sum_{k=1}^{u_n} p_k\right);\\
&\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \overset{\text{iid}}{\sim} N(0, \tau_s); \quad \boldsymbol{\gamma} = \{\gamma_j\}_{j=1}^n \sim N(\mathbf{0}, \tau(\boldsymbol{D}_w - \boldsymbol{M}))
\end{aligned} \tag{3.1}$$

where δ_{θ_k} is the Dirac measure (point mass) located at θ_k and each θ_k, $k = 1, \ldots, K$, is a random sample drawn independently from a base distribution $G_0 = N(0, \tau_s)$ with precision τ_s. Subscripts $u_1, \ldots, u_n$ are indices of θ_k's sampled for n observations. Probability parameters $p_1, \ldots, p_K$ comprises the regular stick breaking weights [Set94] constructed as

$$p_1 = V_1; \quad p_t = V_t \prod_{k<t}(1 - V_k), t = 2, \ldots, K; \quad V_k \overset{\text{iid}}{\sim} Beta(1, \alpha); \tag{3.2}$$

where α stochastically controls the number of distinct values among the K clusters. The total number of DP clusters, K, truncates the stick breaking function. The infinite sum of p_k's is 1. In practice, we simply choose a large enough value for K such that there exists some empty components during the MCMC run. This model ensures that the marginal distribution of $G^{(j)}(w_j)$ for each region j follows an identical DP,

$$G^{(j)}(w_j) = \sum_{k=1}^{K} \sum_{u_1,\ldots,u_i=k,\ldots,u_n} \pi_{u_1,\ldots,u_i=k,\ldots,u_n} \delta_{\theta_{u_1}} \ldots \delta_{\theta_{u_i=k}} \ldots \delta_{\theta_{u_n}} = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}, \tag{3.3}$$

where $\pi_k = P(\sum_{t=1}^{k-1} p_t < F^{(j)}(\gamma_j) < \sum_{t=1}^{k} p_t)$.

We extend this ARDP model to a multivariate ARDP (MARDP) framework for multivariate boundary analysis. Under the MARDP framework, we use the joint version of MDAGAR model illustrated in Section 2.4 to construct spatial components, which avoids

the order selection issue in conditional modelling for multiple diseases. The reminder of this chapter is organized as follows. Section 3.2 develops a MARDP framework for discrete spatial random effects based the hierarchical MDAGAR model for spatial components. Section 3.3 illustrates the decision rule based on FDR for multivariate areal boundary analysis. Section 3.4 presents a simulation study to assess the performance of different hierarchical models in model fitting and difference boundary detection. Section 3.5 introduces a multivariate areal dataset for standardized incidence ratios (SIR) of four cancers in California obtained from the SEER database, and analyzes and discusses difference boundary detection for each cancer as well as across cancers.

3.2 The Multivariate Areally Referenced Dirichlet Process (MARDP)

Let $N = n \times q$ be the total number of observations and recall $\boldsymbol{w} = \left(\boldsymbol{w}_1^\top, \ldots, \boldsymbol{w}_q^\top\right)$ denote the random effects vector, where $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{i,n})^\top$, $i = 1, \ldots, q$. Let $(1,1), \ldots, (1,n), (2,1), \ldots, (2,n), \ldots, (q,1), \ldots, (q,n)$ be the pairwise (i,j) indices corresponding to a vectorized enumeration of the observations $1, \ldots, n, n+1, \ldots, 2n, \ldots, (q-1)n+1, \ldots, N$. Modeling $\boldsymbol{w}$ jointly as an unknown distribution G_N, which itself is modeled as a Dirichlet process (DP), yields the Multivariate Areal DP (MARDP)

$$
\begin{aligned}
&\boldsymbol{w} \sim G_N; G_N|\pi_{u_1,\ldots,u_N}, \boldsymbol{\theta} = \sum_{u_1,\ldots,u_N} \pi_{u_1,\ldots,u_N} \delta_{\theta_{u_1}} \ldots \delta_{\theta_{u_N}};\\
&\pi_{u_1,\ldots,u_N} = Pr\left(\sum_{k=1}^{u_1-1} p_k < F^{(1)}(\gamma_1) < \sum_{k=1}^{u_1} p_k, \ldots, \sum_{k=1}^{u_N-1} p_k < F^{(N)}(\gamma_N) < \sum_{k=1}^{u_N} p_k\right);\\
&\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q\} \sim N_{nq}(\mathbf{0}, \boldsymbol{\Sigma}_\gamma^{-1})
\end{aligned}
\tag{3.4}
$$

where $u_1, \ldots, u_N$ are indices of θ_k's sampled for N the observations with respect to q diseases in n regions, $\boldsymbol{\theta}$ is defined in (3.1) and $p_1, \ldots, p_K$ are specified in (3.2). Spatial components $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2} \ldots, \gamma_{i,n})^\top$ are dependent for each disease i, and $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q$ are mod-

eled jointly by MDAGAR($\rho_1, \ldots, \rho_q, \boldsymbol{\Sigma}$) with covariance matrix $\boldsymbol{\Sigma}_\gamma$ as defined in (2.26), incorporating both associations among the diseases and spatial dependence for each disease. Each $F^{(1)}(\cdot), \ldots, F^{(N)}(\cdot)$ (corresponding to $F^{(1,1)}(\cdot), \ldots, F^{(q,n)}(\cdot)$, respectively) denotes the cumulative distribution functions of the marginal distribution of the corresponding γ_{ij}. Again, each $F^{(i,j)}(\gamma_{ij}) \sim Uniform(0,1)$ but dependence is introduced through the γ_{ij}'s. As a result, spatial random effects $\boldsymbol{w}$ hold all properties that are in univariate ARDP. The marginal distribution for the individual w_{ij} is given as $G^{(i,j)}(w_{ij}) = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}$, where $\pi_k = P\left(\sum_{t=1}^{k-1} p_t < F^{(i,j)}(\gamma_{ij}) < \sum_{t=1}^{k} p_t\right)$. These DPs are dependent across regions as well as diseases with dependent $F^{(i,j)}(\gamma_{ij})$'s and through parameters $p_1, \ldots, p_K$. Hence, the MARDP framework is able to evaluate the difference in w_{ij}'s across diseases. The shared values of θ_k's make it possible to compare effects between diseases.

3.2.1 Model Implementation

We extend (2.4) to a Bayesian hierarchical framework with the posterior distribution

$$p\left(\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{V}, \boldsymbol{\sigma}, \tau_s, \boldsymbol{\rho}, \boldsymbol{A} | \boldsymbol{y}\right) \propto p\left(\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{V}, \boldsymbol{\sigma}, \tau_s, \boldsymbol{\rho}, \boldsymbol{A}\right) \times \prod_{i=1}^{q} \prod_{j=1}^{n} N\left(y_{ij} \,|\, \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_i + w_{ij}, 1/\sigma_i^2\right) \tag{3.5}$$

where $\boldsymbol{V} = (V_1, \ldots, V_K)^\top$. We specify $p\left(\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{V}, \boldsymbol{\sigma}, \tau_s, \boldsymbol{\rho}, \boldsymbol{A}\right)$ as

$$\begin{aligned} &\prod_{k=1}^{K} \left\{N(\theta_k|0,\tau_s) \times Beta(V_k|1,\alpha)\right\} \times \prod_{i=1}^{q} \left\{IG(\sigma_i^2 \,|\, a_e, b_e) \times N(\boldsymbol{\beta}_i \,|\, \mathbf{0}, 1/\sigma_\beta^2 \boldsymbol{I}_{p_i}) \times Unif(\rho_i \,|\, 0,1)\right\} \\ &\quad \times IG(1/\tau_s \,|\, a_s, b_s) \times N\left(\boldsymbol{\gamma} \,|\, \mathbf{0}, \boldsymbol{\Sigma}_\gamma(\boldsymbol{\rho}, \boldsymbol{A})\right) \times IW\left(\boldsymbol{A}\boldsymbol{A}^\top \,|\, \nu, \boldsymbol{R}\right) \times \left|\frac{\partial \boldsymbol{\Sigma}}{\partial a_{ih}}\right| , \end{aligned} \tag{3.6}$$

where $\left|\frac{\partial \boldsymbol{\Sigma}}{\partial a_{ih}}\right|$ is the Jacobian $2^q \prod_{i=1}^{q} a_{ii}^{q-i+1}$ transformation for the prior on $\mathbf{A}\mathbf{A}^\top$ in terms of the Cholesky factor $\mathbf{A}$. We sample the parameters from the posterior distribution in (3.5) using Markov chain Monte Carlo (MCMC) with Gibbs sampling and random walk metropolis

[GL06] implemented in the `R` statistical computing environment. The Appendix 3.7 presents details on the MCMC updating scheme.

3.3 Decision Rule Based on FDR for Selecting Difference Boundaries

Following [LBH15] we formulate difference boundary detection as a multiple comparison problem, where a cancer-specific difference boundary is detected according to the tenability, or not, of $w_{ij} = w_{ij'}$ for $j \sim j'$. To adjust for the multiplicity arising from all pairs of neighbors and, in our case, of diseases as well, a false discovery rate (FDR) is controlled [BH95]. We adopt the Bayesian analogue of FDR [MPR04] in the following manner: We define an edge $(j, j')^i$ as a difference boundary for disease i if the posterior probability $P(w_{ij} \neq w_{ij'}|\boldsymbol{y})$ exceeding a certain threshold t. Denoting $v^i_{(j,j')} = P(w_{ij} \neq w_{ij'}|\boldsymbol{y})$, we define

$$FDR = \frac{\sum_{j\sim j'} I\left(w_{ij} = w_{ij'}\right) I\left(v^i_{(j,j')} > t\right)}{\sum_{j\sim j'} I\left(v^i_{(j,j')} > t\right)},$$

and the estimated FDR is obtained as the posterior expectation

$$\overline{FDR} = \frac{\sum_{j\sim j'} \left(1 - v^i_{(j,j')}\right) I\left(v^i_{(j,j')} > t\right)}{\sum_{j\sim j'} I\left(v^i_{(j,j')} > t\right)}. \tag{3.7}$$

We also compute $\overline{FNR} = \frac{\sum_{j\sim j'} v^i_{(j,j')}\left(1-I\left(v^i_{(j,j')}>t\right)\right)}{m-\sum_{j\sim j'} I\left(v^i_{(j,j')}>t\right)}$ to estimate the False Non-discovery Rate (FNR), where m is the total number of edges (geographic boundaries). In terms of a bivariate loss function $L_{2R} = \left(\overline{FDR}, \overline{FNR}\right)$, the optimal decision minimizes $\overline{FNR}$ subject to $\overline{FDR} \leq \delta$, i.e. the threshold $t = t^\star$ is obtained as [MPR04]:

$$t^\star = \sup\left\{t : \overline{FDR}(t) \leq \delta\right\}. \tag{3.8}$$

The posterior probability $v^i_{(j,j')}$ in (3.7) is defined according to the type of difference boundary. For instance, we use $v^s_{(j,j')} = P\left(w_{ij} \neq w_{ij'}, w_{i'j} \neq w_{i'j'}|\boldsymbol{y}\right)$ for shared boundaries and $v^c_{(j,j')} = P\left(w_{ij} \neq w_{i'j'}, w_{i'j} \neq w_{ij'}|\boldsymbol{y}\right), j < j'$ for mutual cross-disease difference boundaries (i and i' are two different diseases) instead of $v^i_{(j,j')}$ in (3.7).

3.4 Simulation

We present a simulation experiment to compare the performances of MDAGAR and MCAR with two independent-disease models. All models were constructed using the MARDP framework in Section 3.2 and differ only in their specification of $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$.

3.4.1 Data Generation

We generate data over a California county map with 58 counties. We simulated our outcomes y_{ij} using the model in (2.4) with $q = 2$, i.e., two outcomes, and two covariates, $\boldsymbol{x}_{1j} = (1, x_{1j2})^\top$ and $\boldsymbol{x}_{2j} = (1, x_{2j2})^\top$, with $p_1 = p_2 = 2$. We fixed values of x_{1j2} and x_{2j2} by generating them from $N(0, 1)$ independently across regions. The regression slopes were set to $\boldsymbol{\beta}_1 = (2, 5)^\top$ and $\boldsymbol{\beta}_2 = (1, 6)^\top$. We generated values of $\boldsymbol{w} = \left(\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top\right)^\top$ using (3.4) with $K = 15$, $\alpha = 1$, and $\tau_s = 0.25$, while we generated values for $\boldsymbol{\gamma}$ from $N(\boldsymbol{0}, \boldsymbol{\Sigma}_\gamma)$ with $\boldsymbol{\Sigma}_\gamma$ in (2.26) specified by $\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, $\boldsymbol{Q}^{-1}(\rho_i)$ is a spatial autocorrelation matrix with elements $\rho_i^{d(j,j')}$, $\rho_1 = 0.2$ and $\rho_2 = 0.8$, where $d(j, j')$ refers to the distance between the centroids of the jth and j'th counties in California. The specification of $\boldsymbol{A}$ ensures $\text{corr}(\gamma_{1j}, \gamma_{2j}) \approx 0.7$ between the two diseases.

Figure 3.1 shows the map for random effects for disease 1 on the left and disease 2 on the right. There are five different levels in total for both diseases with values -2.67, -1.73, -0.98, 0.42 and 0.77 ordered from the smallest to largest. As a result, we found 75 "true difference boundaries" delineating clusters with substantially different values for disease 1

and 78 "true difference boundaries" for disease 2. Moreover, there are 77 cross-disease difference boundaries delineating random effects for disease 1 from disease 2 in neighboring regions, i.e. $w_{1j} \neq w_{2j'}, j \sim j'$, and $j < j'$; there are 95 cross-disease difference boundaries separating disease 2 from disease 1 in the neighboring regions, i.e. $w_{2j} \neq w_{1j'}, j \sim j'$, and $j < j'$.

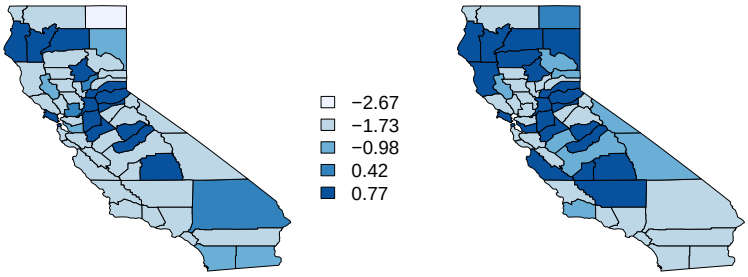


Figure 3.1: A map of the simulated data for random effects for disease 1 (left) and disease 2 (right) showing five different levels, each with its own value. There are 75 boundary segments that separate regions for disease 1 and 78 difference boundaries for disease 2.

3.4.2 Model Comparison

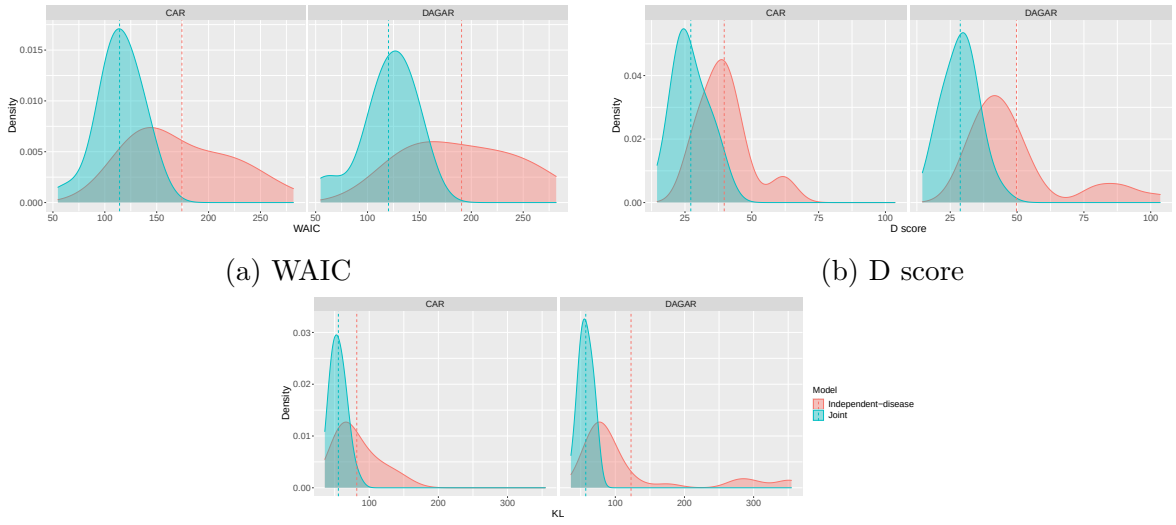
Fixing the values of $\boldsymbol{w}$ generated as above, we simulated 30 datasets for the outcome $y_{ij} \sim N\left(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}_i + \phi_{ij}, 1/sigma_i^2\right)$, where $\sigma_1^2 = \sigma_2^2 = 0.1$. We analyzed the 30 replicated datasets using (3.5) with vague priors specified in (3.6) as $a_s = 2$, $b_s = 0.1$, $a_e = 2$, $b_e = 0.1$, $\sigma_\beta^2 = 1000$, $\alpha = 1$, $\nu = 2$ and $\boldsymbol{R} = \text{diag}(0.1, 0.1)$. The same set of priors were used for both MDAGAR and MCAR as they have the same number of parameters with similar interpretations. The joint multivariate settings were compared with corresponding independent-disease models for CAR and DAGAR respectively. For independent-disease models, spatial components are assumed to be independent between diseases. Hence $\boldsymbol{A} = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix}$

and $\boldsymbol{\Sigma}_\gamma = \begin{pmatrix} a_{11}^2 \boldsymbol{Q}^{-1}(\rho_1) & \boldsymbol{O} \\ \boldsymbol{O} & a_{22}^2 \boldsymbol{Q}^{-1}(\rho_2) \end{pmatrix}$ is block diagonal. We refer to the independent-disease models by DAGAR$_{ind}$ and CAR$_{ind}$ according to whether $\boldsymbol{Q}(\rho_i)$ is specified by DAGAR and CAR, respectively. We used the same priors as for the joint models except for $\boldsymbol{A}$, which is now specified by $a_{ii}^2 \sim IG(a_v, b_v), a_v = 2, b_v = 0.1$ for $i = 1, 2$. All models were executed in the `R` statistical computing environment and inference was obtained from 5000×2 (chains) $= 10000$ MCMC samples from (3.5) for each model.

We compared MDAGAR, MCAR, DAGAR$_{ind}$ and CAR$_{ind}$ using the Widely Applicable Information Criterion (WAIC) [Wat10, GHV14] and a predictive loss criterion based on a balanced loss function for replicated data sets [GG98]. For the latter, we drew replicates $y_{\text{rep},ij}^{(\ell)} \sim N\left(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_i^{(\ell)} + w_{ij}^{(\ell)}, 1/\sigma_i^{2(\ell)}\right)$ for each posterior sample $\ell = 1, \ldots, L$ and computed $D = G + P$, where $G = \sum_{i=1}^q \sum_{j=1}^n (y_{ij} - \bar{y}_{\text{rep},ij})^2$ and $P = \sum_{i=1}^q \sum_{j=1}^n \sigma_{\text{rep},ij}^2$, $\bar{y}_{\text{rep},ij} = \dfrac{1}{L}\sum_{\ell=1}^{L} y_{\text{rep},ij}^{(\ell)}$, and $\sigma_{\text{rep},ij}^2$ is the variance of $y_{\text{rep},ij}^{(\ell)}$ for $\ell = 1, \ldots, L$. Both WAIC and D reward goodness of fit and penalize model complexity. Figure 3.2 plots values of WAICs (3.2a) and D scores (3.2b) over the 30 data sets for the four models. Compared with the two independent-disease models, the two joint models exhibit much better performance with lower WAIC and D scores. This, unsurprisingly, indicates the benefits of capturing dependence among diseases in terms of model choice. MDAGAR and MCAR perform comparably, although CAR$_{ind}$ seems to be slightly preferred to DAGAR$_{ind}$.

We also computed the Kullback-Leibler Divergence, $D_{KL}\left(p(\boldsymbol{y}_{true}) || p(\boldsymbol{y})\right)$, between the true density $p(\boldsymbol{y}_{true})$ and the four models. Here, $p(\boldsymbol{y}) = N(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\phi}, diag(\boldsymbol{\sigma}) \otimes \boldsymbol{I}_n)$ is the density from each candidate model and $p(\boldsymbol{y}_{true}) = N\left(\boldsymbol{y}_{true} \mid \boldsymbol{X}\boldsymbol{\beta}_{true} + \boldsymbol{\phi}_{true}, diag(\boldsymbol{\sigma}_{true}) \otimes \boldsymbol{I}_n\right)$, where diag($\boldsymbol{\sigma}$) is a diagonal matrix with σ_i^2 as i-th diagonal element, and $\boldsymbol{X}$ is a block diagonal design matrix with $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{in})^\top$ as diagonal blocks. Since $D_{KL}\left(p(\boldsymbol{y}_{true}) || p(\boldsymbol{y})\right)$ is a function of the model parameters, we can compute its posterior distribution given each data set. We collect the posterior means from each dataset and plot them using a density-

smoother in Figure 3.2c for the four models. These plots clearly show that the joint models, MCAR and MDAGAR, have smaller KL divergences from the true model than have CAR$_{ind}$ and DAGAR$_{ind}$. We also evaluated parameter estimates from the four models as discussed in Appendix 3.7.2.



(a) WAIC

(b) D score

(c) KL

Figure 3.2: Density plots for WAICs, D scores and mean $D_{KL}(p(\boldsymbol{y}_{true})||p(\boldsymbol{y}))$ over 30 datasets as shown in (a), (b) and (c) respectively, using two joint models, MCAR (blue plot in CAR panel) and MDAGAR (blue plot in DAGAR panel), and two independent-disease models, CAR$_{ind}$ (red plot in CAR panel) and DAGAR$_{ind}$ (red plot in DAGAR panel). The dotted vertical line shows the mean for each plot.

Turning to boundary detection, we computed $P(w_{ij} \neq w_{i'j'}|\boldsymbol{y})$ for $i, i' = 1, 2$ and for every pair of neighboring regions (j, j'). Given these posterior probabilities, we obtained the corresponding boundary detection results (sensitivity and specificity) between and across diseases over our 30 simulated datasets using our four models. Table 3.1 presents these results. Given the true number of difference boundaries, sensitivities and specificities were calculated by choosing difference boundaries as a fixed number T of edges ranked in terms of highest posterior probabilities. This was repeated for $T = 60, 65, 70, 75, 80, 85$ for disease 1, disease 2 and disease 1 vs. 2, while $T = 70, 75, 80, 85, 90, 95$ were used for disease 2 vs. 1. Overall, the two joint models produce comparable detection results and outperform the

two independent-disease models in terms of sensitivity and specificity under all scenarios. In most scenarios, MDAGAR appears to outperform MCAR in terms of specificity and sensitivity for disease 1, which may be attributed to the better estimation of autocorrelation parameters in DAGAR when ρ is small ($\rho_1 = 0.2$) [DBH19]. By choosing T close to the true number of difference boundaries for each disease scenario, MDAGAR and MCAR are able to detect about 85% of the true boundaries with specificity and sensitivity both around 85% for disease 1, disease 2 and disease 1 vs. 2. When comparing diseases 2 vs. 1, MDAGAR and MCAR detect about 82% of the true boundaries with specificity and sensitivity around 82% when $T = 85$. In most of these scenarios, the disease-independent models are more likely to be false positive, recognizing the null case (i.e. $w_{ij} = w_{i'j'}$) as difference boundaries.

Table 3.1: Boundary detection results (sensitivity and specificity) in the simulation study (30 datasets generated on the California map) within each disease and across two diseases using MCAR, MDAGAR, CAR_{ind} and DAGAR_{ind} methods.

		Disease 1		Disease 2		Disease 1 vs 2				Disease 2 vs 1	
T	Methods	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	T	Methods	Specificity	Sensitivity
60	MDAGAR	0.951	0.781	0.953	0.744	0.917	0.774	70	MDAGAR	0.920	0.737
	MCAR	0.924	0.783	0.935	0.751	0.928	0.756		MCAR	0.913	0.721
	DAGAR_{ind}	0.945	0.767	0.956	0.753	0.902	0.732		DAGAR_{ind}	0.900	0.726
	CAR_{ind}	0.909	0.763	0.909	0.732	0.902	0.740		CAR_{ind}	0.899	0.714
65	MDAGAR	0.923	0.816	0.933	0.786	0.891	0.803	75	MDAGAR	0.895	0.767
	MCAR	0.900	0.814	0.918	0.791	0.903	0.788		MCAR	0.887	0.752
	DAGAR_{ind}	0.874	0.799	0.890	0.793	0.833	0.770		DAGAR_{ind}	0.858	0.758
	CAR_{ind}	0.885	0.796	0.888	0.759	0.879	0.771		CAR_{ind}	0.877	0.746
70	MDAGAR	0.886	0.846	0.898	0.818	0.854	0.828	80	MDAGAR	0.858	0.794
	MCAR	0.870	0.845	0.884	0.822	0.876	0.824		MCAR	0.857	0.784
	DAGAR_{ind}	0.797	0.816	0.842	0.816	0.763	0.800		DAGAR_{ind}	0.780	0.799
	CAR_{ind}	0.810	0.822	0.858	0.790	0.834	0.804		CAR_{ind}	0.824	0.781
75	MDAGAR	0.839	0.869	0.854	0.847	0.812	0.851	85	MDAGAR	0.816	0.820
	MCAR	0.826	0.864	0.849	0.852	0.838	0.850		MCAR	0.823	0.816
	DAGAR_{ind}	0.765	0.832	0.785	0.835	0.718	0.820		DAGAR_{ind}	0.705	0.826
	CAR_{ind}	0.733	0.848	0.796	0.823	0.761	0.835		CAR_{ind}	0.760	0.816
80	MDAGAR	0.781	0.884	0.803	0.872	0.767	0.871	90	MDAGAR	0.766	0.845
	MCAR	0.776	0.882	0.804	0.876	0.785	0.871		MCAR	0.783	0.849
	DAGAR_{ind}	0.690	0.859	0.727	0.861	0.666	0.844		DAGAR_{ind}	0.672	0.844
	CAR_{ind}	0.670	0.865	0.707	0.857	0.694	0.853		CAR_{ind}	0.698	0.844
85	MDAGAR	0.701	0.904	0.749	0.893	0.715	0.890	95	MDAGAR	0.710	0.871
	MCAR	0.720	0.899	0.750	0.895	0.728	0.886		MCAR	0.728	0.874
	DAGAR_{ind}	0.622	0.881	0.694	0.877	0.632	0.860		DAGAR_{ind}	0.603	0.866
	CAR_{ind}	0.597	0.880	0.646	0.880	0.605	0.874		CAR_{ind}	0.614	0.868

Note: The first column "T" is the number of edges fixed as difference boundaries in terms of highest posterior probabilities.

Our methodology also allows us to detect "disease differences" within the same county, i.e. $P(w_{1j} \neq w_{2j'} \mid \boldsymbol{y})$. This reflects difference in the random effects between two diseases in the same county. There are 20 counties with true "disease difference" in Figure 3.1. Ta-

ble 3.2 shows sensitivity and specificity for detecting "disease difference" in the same county using the four models over 30 datasets. We chose $T = 15, 20, 22, 25, 30$ regions with the highest posterior probabilities as differences between diseases. Unsurprisingly, MDAGAR and MCAR again excel over the two independent-disease models in all scenarios with a resulting sensitivity and specificity of about 80% when $T = 22$. Moreover, DAGAR models tend to have better detection than CAR models indicating better interpretation of the association among diseases.

Table 3.2: Sensitivity and specificity in the simulation study (30 datasets generated on the California map) for "disease difference" in the same region using MCAR, MDAGAR, CAR_{ind} and DAGAR_{ind} methods.

T	Methods	Specificity	Sensitivity	T	Methods	Specificity	Sensitivity
15	MDAGAR	0.914	0.690	20	MDAGAR	0.844	0.773
	MCAR	0.900	0.672		MCAR	0.821	0.742
	DAGAR_{ind}	0.892	0.610		DAGAR_{ind}	0.790	0.715
	CAR_{ind}	0.889	0.593		CAR_{ind}	0.798	0.675
22	MDAGAR	0.807	0.795	25	MDAGAR	0.746	0.817
	MCAR	0.760	0.772		MCAR	0.707	0.802
	DAGAR_{ind}	0.722	0.762		DAGAR_{ind}	0.680	0.787
	CAR_{ind}	0.761	0.693		CAR_{ind}	0.649	0.743
30	MDAGAR	0.646	0.857				
	MCAR	0.617	0.855				
	DAGAR_{ind}	0.594	0.845				
	CAR_{ind}	0.545	0.800				

Note: The first column "T" is the number of edges fixed as difference boundaries in terms of highest posterior probabilities.

3.5 Analysis of SEER Dataset with Four Cancers

3.5.1 Data Example

For the incidence of the four cancers: lung, esophageal, larynx and colorectal cancer, we analyze a dataset including the observed counts of incidence (Y_{ij}) for each cancer $i = 1, 2, 3, 4$

in each county $j = 1, 2, \ldots, 58$ of California between 2012 and 2016. Given the population N_j in county j, we calculated the expected number of cases, $E_{ij} = \frac{\sum_{j=1}^{58} Y_{ij}}{\sum_{j=1}^{58} N_j} N_j$, and plotted the standardized incidence ratios ($\text{SIR}_{ij} = Y_{ij}/E_{ij}$) on a California map showing the 58 counties for the four cancers, as shown in Figure 3.3. Cutoffs for the different levels of SIRs are quintiles for each cancer.

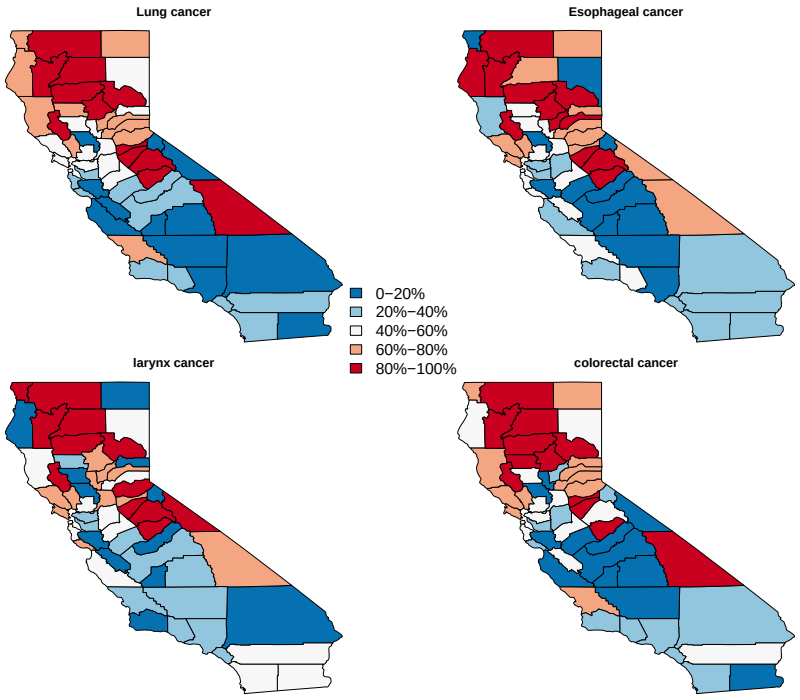


Figure 3.3: Maps of standardized incidence ratios (SIR) for lung, esophageal, larynx and colorectal cancer in California, $2012 - 2016$.

As an exploratory tool to assess associations among the cancers, we calculated Pearson's correlation for each pair of cancers by regarding SIRs in different counties as independent samples and found that the correlation coefficients between the incidence ratios for all four cancers to be relatively high with values $0.5 - 0.9$. Next, to explore the spatial association for each cancer, we calculated Moran's I based upon the rth order neighbors for each cancer and plotted the areal correlogram [BCG14]. Defining distance intervals as $(0, d_1], (d_1, d_2], (d_2, d_3], \ldots$, the rth order neighbors refer to units with distance in $(d_{r-1}, d_r]$,

i.e. within distance d_r but separated by more than d_{r-1}. The distance is the Euclidean distance from an Albers map projection of California. Figure 3.4 reveals that spatial associations in lung, esophageal and colorectal cancers clearly diminish with increasing r, although the pattern is less pronounced for larynx.

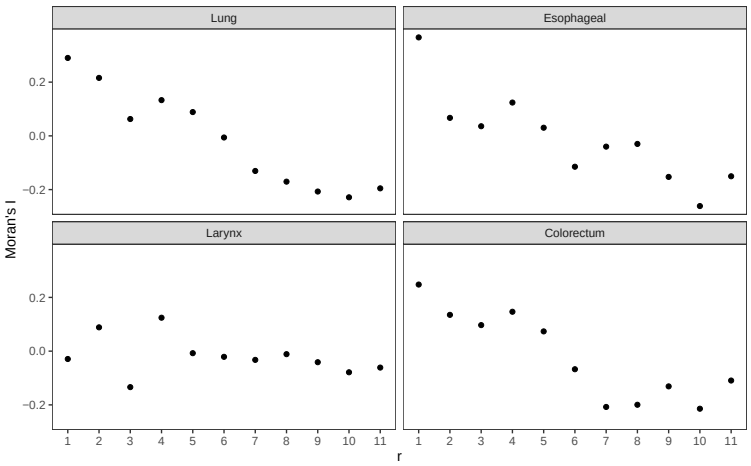


Figure 3.4: Moran's I of rth order neighbors for lung, esophageal, larynx and colorectal cancer.

For an insight into difference boundaries for each cancer, we calculated the difference in SIR between each pair of neighboring counties (139 pairs in total), i.e., $|SIR_{ij} - SIR_{ij'}|, j \sim j'$. By ranking the differences from largest to smallest, we selected the first 70 pairs (half of the total pairs) with the largest differences as the difference boundaries for each cancer as shown in Figure 3.5. The four cancers exhibit similar patterns in boundary detection that more boundaries are detected in the north and the borders of California. Counties along the central corridor of California, ranging from central to south, tend to be in the same cluster.

3.5.2 Data Analysis

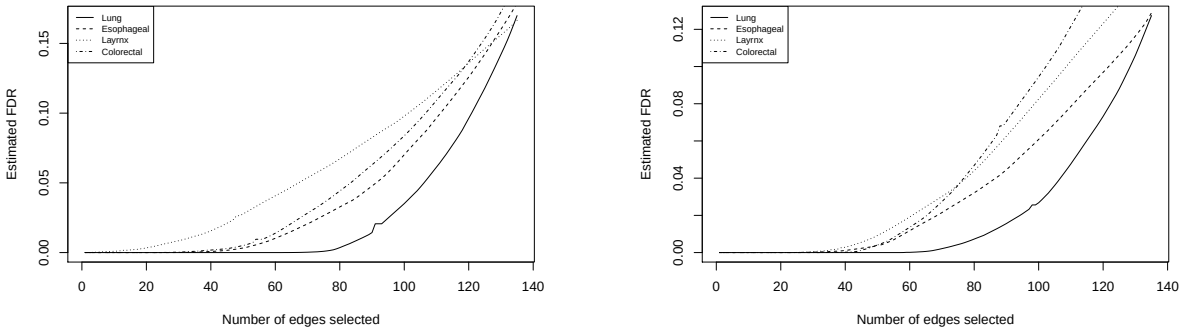
We analyzed the dataset mentioned in Section 3.5.1 using a Poisson spatial regression model, i.e. $Y_{ij} \overset{\text{ind}}{\sim} Poisson\left(E_{ij} \exp\left(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}_i + w_{ij}\right)\right)$, $i = 1, \ldots, 4$, $j = 1, \ldots, 58$. Applying prior specification as in the simulation study, we implemented MARDP as discussed in Section 3.2.1

Figure 3.5: Boundaries (in red) selected as the first 70 pairs with largest differences for lung, esophageal, larynx and colorectal cancer.

using MDAGAR and MCAR. Posterior inference is based upon 10000 MCMC samples after 20000 iterations of burn-in for diagosing convergence.

Without accounting for covariates, we detected difference boundaries for SIR of each cancer and across cancers. First, regarding boundary detection for each cancer, we set up a threshold to control for FDR as in (3.8). Figure 3.6 plots the change of estimated FDR with different numbers of edges selected as difference boundaries for the four cancers individually using MDAGAR (3.6a) and MCAR (3.6b). In general, MDAGAR and MCAR render similar trends in FDR curves, which are close to each other for esophageal, colorectal and larynx cancers while lung cancer exhibits much smaller values. With MDAGAR the FDR increases slightly faster for larynx. Apparently, under the same threshold value we will detect more boundaries for lung and fewer boundaries for larynx cancer. Setting $\delta = 0.025$ in (3.8), Figure 3.7 shows difference boundaries (highlighted in red) detected by MDAGAR and MCAR in SIR maps for the four cancers. Maps from MDAGAR and MCAR are consistent with each other and the number of difference boundaries detected by the two models are also

similar for each cancer, albeit with fewer boundaries (47 edges with posterior probabilities above the threshold $t^\star$ in (3.8)) detected for larynx under MDAGAR. For lung cancer 95 boundaries are detected in total, which is considerably higher than the other three cancers.

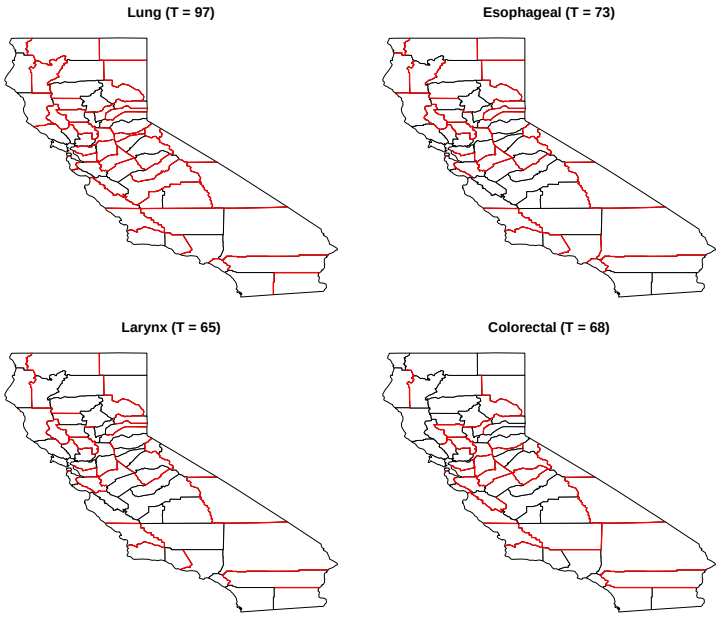


(a) MDAGAR (b) MCAR

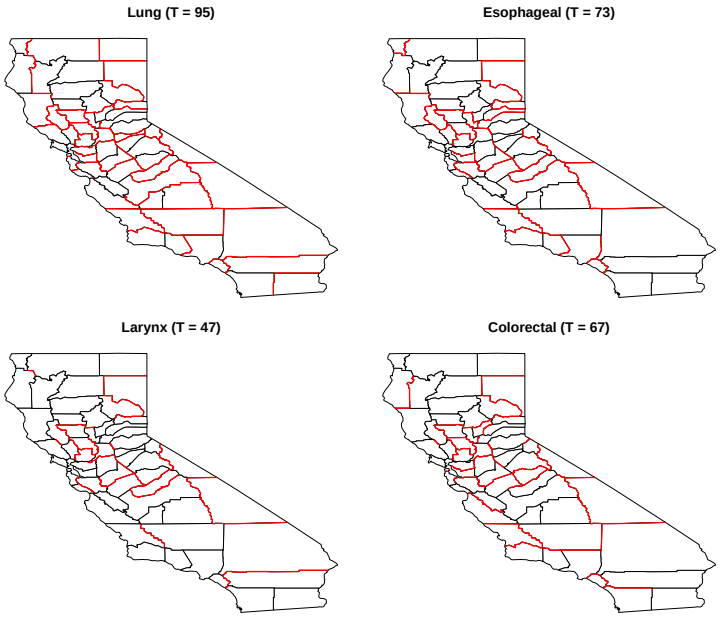
Figure 3.6: Estimated FDR curves plotted against the number of selected difference boundaries for four cancers using MDAGAR and MCAR.

Table 3.3 provides an exhaustive list of the cancer boundaries detected by MDAGAR in Figure 3.7. This "lookup table" contains the names of adjacent counties ranked in decreasing order of $P(w_{ij} \neq w_{ij'} \,|\, \boldsymbol{y})$ for the four cancers, offering a detailed reference for health administrators to identify substantial spatial health barriers. Around $80\% - 90\%$ of the boundaries listed here are also detected by MCAR. For each cancer, we see some clusters and islands (regions fully encompassed by difference boundaries with all neighbors within California) in the map. For example, the northern counties of Siskiyou, Shasta, Tehama, Glenn, Butte, Plumas and Trinity appear to form a cluster for all cancers. Similarly, the central and southern counties of Santa Clara, Merced, San Benito, Fresno, Kings, Tulare and Kern appear in the same cluster for esophageal, larynx and colorectal cancers. Lassen and Inyo are islands (for all cancers) with substantially smaller and larger effects, respectively, than their neighbors. San Luis Obispo is an island with larger effects for lung, esophageal and colorectal cancers, while Lake is an island with larger effects for lung and esophageal cancers. A California map with county names labeled is shown in Appendix Figure 3.12 for

reference.



(a) MDAGAR



(b) MCAR

Figure 3.7: Difference boundaries (highlighted in red) detected by (a) MDAGAR and (b) MCAR in SIR map for four cancers individually when $\delta = 0.025$. The values in brackets are the number of difference boundaries detected.

Table 3.3: Names of adjacent counties that have significant boundary effects from the MDA-GAR model for each cancer when $\delta = 0.025$. The numbers in the first column are ranked according to $P(\phi_{id} \neq \phi_{jd}|\boldsymbol{y})$. Note: Number $1-61$ for lung cancer, $1-26$ for esophageal cancer and $1-22$ for colorectal cancer are ranked by initial letters with $P(\phi_{id} \neq \phi_{jd}|\boldsymbol{y}) = 1$.

Rank	Lung (95)	Esophageal (73)	Layrnx (47)	Colorectal (67)
1	Alameda, Contra costa	Butte, Sutter	Merced, Tuolumne	Alameda, Contra costa
2	Alameda, San joaquin	Calaveras, San joaquin	Mariposa, Merced	Amador, San joaquin
3	Alameda, Santa clara	Calaveras, Stanislaus	Napa, Yolo	Butte, Sutter
4	Amador, El dorado	El dorado, Sacramento	Lake, Yolo	Calaveras, San joaquin
5	Amador, Sacramento	Fresno, Inyo	San joaquin, Santa clara	Contra costa, San joaquin
6	Amador, San joaquin	Inyo, Kern	San mateo, Santa clara	Fresno, Inyo
7	Butte, Colusa	Inyo, San bernardino	Inyo, Kern	Inyo, Kern
8	Butte, Sutter	Inyo, Tulare	Madera, Tuolumne	Inyo, Tulare
9	Calaveras, San joaquin	Kern, San luis obispo	Sacramento, Yolo	Kern, Los angeles
10	Calaveras, Stanislaus	Kings, San luis obispo	Fresno, Inyo	Kern, San bernardino
11	Colusa, Glenn	Lake, Mendocino	Calaveras, San joaquin	Kern, San luis obispo
12	Colusa, Lake	Lake, Napa	Inyo, San bernardino	Kern, Santa barbara
13	Contra costa, Sacramento	Lake, Yolo	Calaveras, Stanislaus	Kern, Ventura
14	Contra costa, Solano	Los angeles, Ventura	Inyo, Tulare	Kings, San luis obispo
15	El dorado, Sacramento	Madera, Mariposa	Santa clara, Stanislaus	Lake, Yolo
16	Fresno, Inyo	Madera, Tuolumne	Orange, San diego	Madera, Mariposa
17	Fresno, Madera	Mariposa, Merced	Madera, Mariposa	Mariposa, Merced
18	Fresno, Tulare	Merced, Tuolumne	Orange, Riverside	Merced, Tuolumne
19	Glenn, Lake	Mono, Tuolumne	Contra costa, Sacramento	Monterey, San luis obispo
20	Imperial, Riverside	Monterey, San luis obispo	Napa, Solano	Napa, Yolo
21	Imperial, San diego	Napa, Yolo	Lassen, Plumas	San luis obispo, Santa barbara
22	Inyo, Kern	Orange, Riverside	Kern, San luis obispo	Solano, Yolo
23	Inyo, Mono	Orange, San diego	Lassen, Shasta	Sacramento, Yolo
24	Inyo, San bernardino	San joaquin, Santa clara	Merced, Stanislaus	Inyo, San bernardino
25	Inyo, Tulare	San mateo, Santa clara	El dorado, Sacramento	Lassen, Shasta
26	Kern, San luis obispo	Stanislaus, Tuolumne	Colusa, Glenn	Santa clara, Stanislaus
27	Kern, Santa barbara	Napa, Solano	Butte, Colusa	Monterey, Santa cruz
28	Kern, Ventura	Merced, Stanislaus	Santa clara, Santa cruz	San mateo, Santa clara
29	Kings, San luis obispo	Kern, Santa barbara	Lassen, Modoc	Calaveras, Stanislaus
30	Lake, Mendocino	Lassen, Shasta	Kings, San luis obispo	Napa, Solano
31	Lake, Napa	Santa clara, Stanislaus	Stanislaus, Tuolumne	Butte, Yuba
32	Lake, Sonoma	Kern, Ventura	Mono, Tuolumne	Colusa, Lake
33	Lake, Yolo	Lassen, Plumas	Mariposa, Stanislaus	Orange, Riverside
34	Lassen, Plumas	Butte, Colusa	Sutter, Yolo	Colusa, Glenn
35	Lassen, Shasta	Sutter, Yolo	Lake, Sonoma	Madera, Tuolumne
36	Los angeles, Orange	Colusa, Glenn	Humboldt, Siskiyou	Sacramento, San joaquin
37	Los angeles, Ventura	Solano, Yolo	Colusa, Lake	San joaquin, Santa clara
38	Madera, Mariposa	Placer, Sacramento	Inyo, Mono	Alameda, Santa clara
39	Madera, Tuolumne	Lake, Sonoma	Amador, San joaquin	Inyo, Mono
40	Mariposa, Merced	Mariposa, Stanislaus	Butte, Sutter	Mariposa, Stanislaus
41	Mariposa, Stanislaus	Los angeles, Orange	Plumas, Sierra	San francisco, San mateo
42	Merced, Stanislaus	Sacramento, Yolo	Riverside, San bernardino	El dorado, Sacramento
43	Merced, Tuolumne	Colusa, Lake	Solano, Yolo	Merced, Stanislaus
44	Mono, Tuolumne	Inyo, Mono	Sacramento, San joaquin	Amador, Sacramento
45	Monterey, San luis obispo	Lassen, Modoc	Alpine, Mono	Humboldt, Mendocino
46	Napa, Solano	Mendocino, Trinity	Madera, Merced	Butte, Colusa
47	Napa, Yolo	Amador, San joaquin	Fresno, Madera	Placer, Sutter
48	Orange, Riverside	Alameda, Santa clara		Colusa, Yolo
49	Orange, San bernardino	Madera, Merced		Lassen, Plumas

50	Orange, San diego	Santa clara, Santa cruz	Lassen, Modoc
51	Riverside, San bernardino	Alpine, Tuolumne	Plumas, Yuba
52	Sacramento, San joaquin	Mendocino, Tehama	Monterey, San benito
53	Sacramento, Yolo	San luis obispo, Santa barbara	Riverside, San diego
54	San francisco, San mateo	Nevada, Sierra	Los angeles, Orange
55	San joaquin, Santa clara	Orange, San bernardino	Orange, San bernardino
56	San luis obispo, Santa barbara	Glenn, Lake	Orange, San diego
57	San mateo, Santa clara	Fresno, Madera	Sutter, Yolo
58	Santa clara, Stanislaus	Fresno, Mono	Humboldt, Trinity
59	Solano, Yolo	Alpine, Amador	Alpine, Mono
60	Stanislaus, Tuolumne	Humboldt, Mendocino	Mono, Tuolumne
61	Sutter, Yolo	Placer, Sutter	Fresno, Mono
62	Amador, Calaveras	Kern, Monterey	Plumas, Sierra
63	Alpine, Amador	Sierra, Yuba	Lake, Sonoma
64	Humboldt, Trinity	Alameda, Contra costa	Nevada, Yuba
65	Mendocino, Trinity	Kings, Monterey	Madera, Merced
66	Butte, Yuba	Plumas, Sierra	Alpine, El dorado
67	Madera, Merced	Sutter, Yuba	Mariposa, Tuolumne
68	Fresno, Kings	Alameda, San joaquin	
69	Lassen, Modoc	Los angeles, San bernardino	
70	Mendocino, Sonoma	Nevada, Placer	
71	Placer, Sacramento	Del norte, Siskiyou	
72	Alameda, Stanislaus	Alpine, Calaveras	
73	San mateo, Santa cruz	Del norte, Humboldt	
74	Fresno, Monterey		
75	Mendocino, Tehama		
76	Humboldt, Siskiyou		
77	Plumas, Sierra		
78	Nevada, Sierra		
79	Alpine, Calaveras		
80	Fresno, Mono		
81	Colusa, Sutter		
82	Placer, Sutter		
83	Alpine, Mono		
84	Modoc, Shasta		
85	Modoc, Siskiyou		
86	Del norte, Siskiyou		
87	Alpine, Tuolumne		
88	Sierra, Yuba		
89	Colusa, Yolo		
90	Plumas, Yuba		
91	Kern, Los angeles		
92	Kern, San bernardino		
93	Kern, Tulare		
94	Kern, Kings		
95	San benito, Santa cruz		

For difference boundaries between cancers, we considered the shared difference boundaries and cross-cancer boundaries. Here, we only show results from MDAGAR (MCAR is similar). The shared difference boundaries are defined as common boundaries detected for different

cancers. Figure 3.8 exhibits the shared boundaries for each pair of cancers, i.e. $P(w_{ij} \neq w_{ij'}, w_{i'j} \neq w_{i'j'}|\boldsymbol{y}), i \neq i'$. Consistent with results for individual cancers in Figure 3.7, Lassen and Inyo are islands with shared difference boundaries within California for all pairs of cancers. San Luis Obispo is a shared island for [lung, esophageal], [lung, colorectal] and [esophageal, colorectal], while Lake is a shared island only for [lung, esophageal]. For cross-cancer difference boundaries, we define a mutual cross-cancer boundary from $P(w_{ij} \neq w_{i'j'}, w_{i'j} \neq w_{ij'}|\boldsymbol{y}), j \sim j', j < j'$, which separates effects for different cancers mutually in neighboring counties (see Figure 3.9). In conjunction with Figure 3.7, we observe that the shared difference boundaries for each pair of cancers also tend to be mutual cross-cancer difference boundaries for the same pair. This indicates high correlation between the SIR's for different cancers. This is consistent with the estimated average correlation of $0.7 - 0.9$ between cancers in the same region.

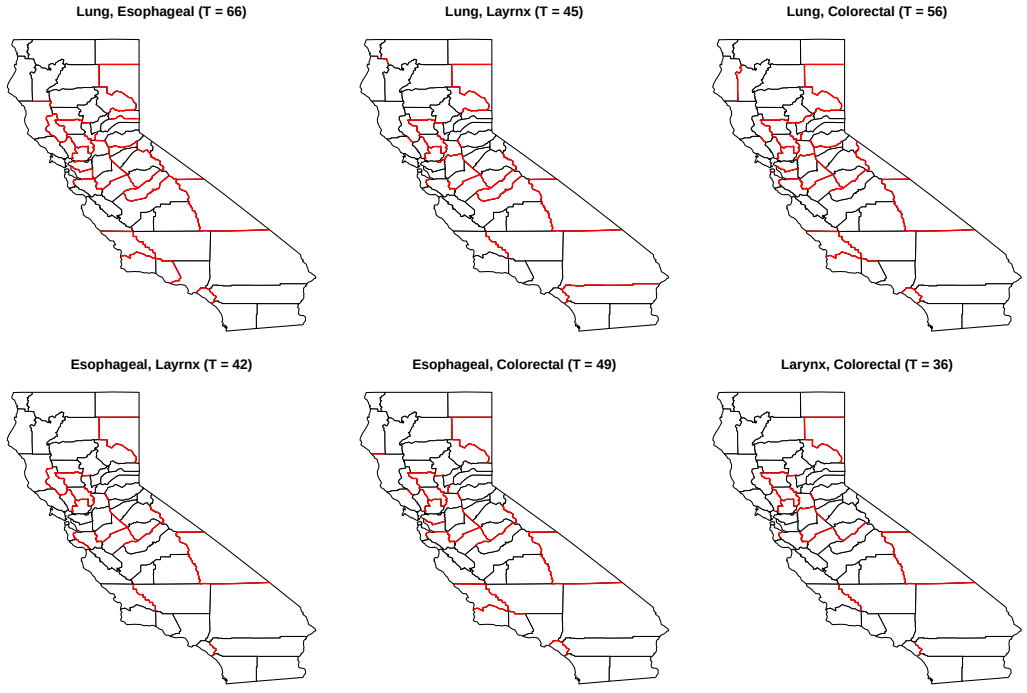


Figure 3.8: Shared difference boundaries (highlighted in red) detected by MDAGAR for each pair of cancers in SIR map when $\delta = 0.025$. The values in brackets are the number of difference boundaries detected.

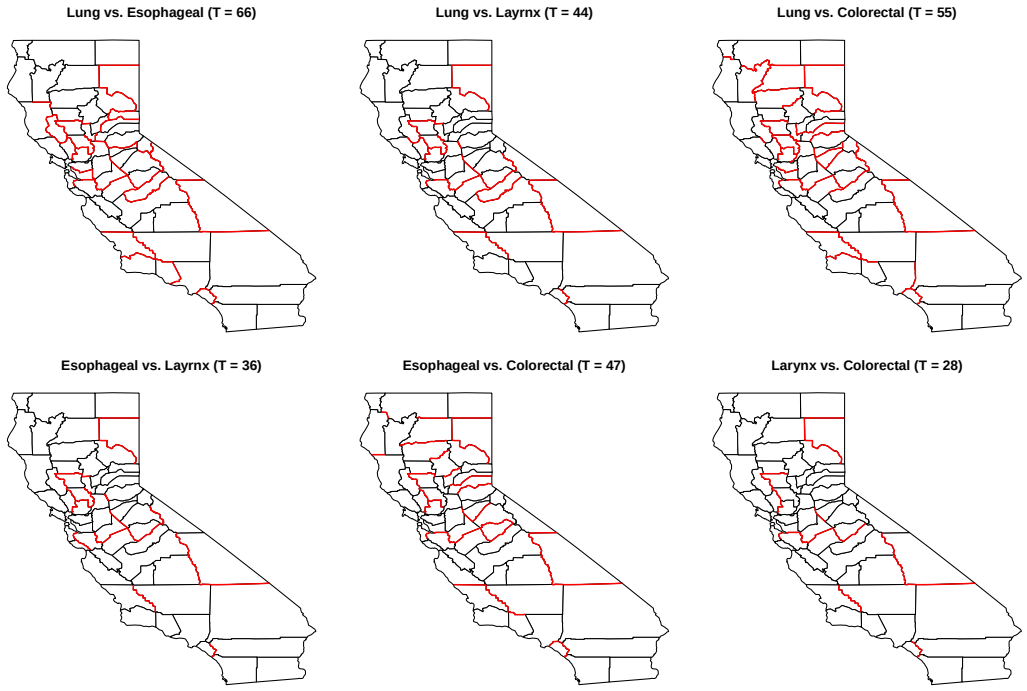


Figure 3.9: Mutual cross-cancer difference boundaries (highlighted in red) detected by MDAGAR for each pair of cancers in SIR map when $\delta = 0.025$. The values in brackets are the number of difference boundaries detected.

We also compare the two joint models with the two independent models. Table 3.4 presents the predictive loss criterion D score for the models. For Poisson regression, replicates for each data point are replaced by $y^{(\ell)}_{\text{rep},ij} = Y^{(\ell)}_{\text{rep},ij}/E_{ij}$, where $Y^{(\ell)}_{\text{rep},ij} \sim Poisson$ $\left(E_{ij} \exp\left(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}_i^{(\ell)} + w_{ij}^{(\ell)}\right)\right)$. The D scores are calculated for each cancer and added up for the four cancers to produce D_{sum}. Unsurprisingly, MDAGAR and MCAR are very comparable (MCAR having a slightly lower score). Both models clearly excel over the two independent models according to D_{sum}, with prominent contributions from lower D scores in lung, esophageal and larynx cancers. DAGAR$_{ind}$ and CAR$_{ind}$ detect fewer difference boundaries for each cancer under the same FDR threshold compared with MDAGAR and MCAR. When $\delta = 0.1$, DAGAR$_{ind}$ and CAR$_{ind}$ produce similar patterns with a similar number of boundaries as detected by MDAGAR and MCAR with $\delta = 0.025$ for lung, esophageal and colorectal cancer (see Figure 3.7); fewer boundaries are detected for larynx. Detecting the

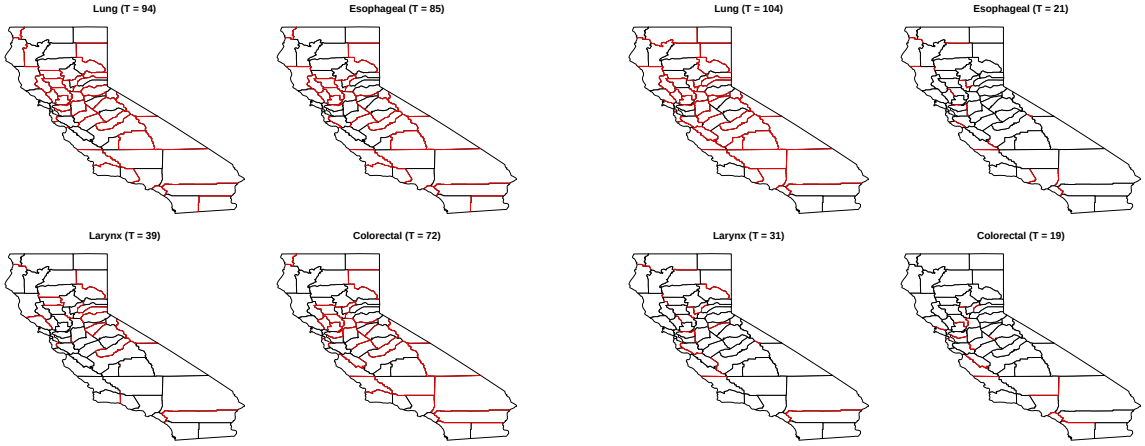
shared boundaries between the three cancers pairwise using DAGAR$_{ind}$ and CAR$_{ind}$ under the same setting ($\delta = 0.1$) reveals fewer shared boundaries.

Table 3.4: Predictive loss criterion D score under four models: MDAGAR, MCAR, DAGAR$_{ind}$, CAR$_{ind}$ using SEER dataset. The D scores are calculated for each cancer individually and added up to D_{sum} for all cancers.

Models	D_{lung}	$D_{esophageal}$	D_{larynx}	$D_{colorectal}$	D_{sum}
MDAGAR	2.49	26.72	45.20	2.27	76.68
MCAR	2.31	26.12	44.17	2.08	74.67
DAGAR$_{ind}$	4.91	30.36	49.28	2.05	86.61
CAR$_{ind}$	5.65	32.85	47.35	2.40	88.26

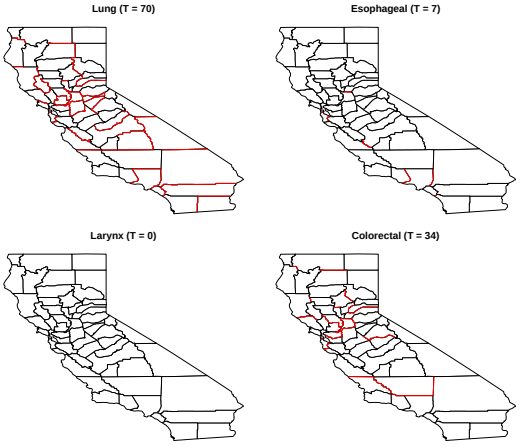
We explore the impact of risk factors in boundary detection by including a potential common risk factor for cancers, adult smoking rates (smoking$_{ij}$), for 2014–2016 obtained from the California Tobacco Facts and Figures 2018 database [Cal18a], and two county attributes that possibly affect the SIR: percentages of residents older than 65 years old (age$_{ij}$) and unemployed residents (unemployed$_{ij}$). Both county attributes are common for different cancers and extracted from the SEER*Stat database [Nat19] for the same period, 2012–2016. Maps of these three covariates are shown in Figure 2.12 using quintiles as cutoffs. Adding the three covariates sequentially, Figure 3.10 shows difference boundaries detected by MDAGAR after accounting for only "smoking" in Figure 3.10a; accounting for "smoking" and "age" in Figure 3.10b; and accounting for all three covariates ("smoking", "age" and "unemployed") in Figure 3.10c for all four cancers when $\delta = 0.025$. Table 3.5 shows posterior means (95% credible intervals) for regression coefficients and autocorrelation parameters estimated without any of the covariates (only an intercept). Unsurprisingly, regression slopes for the percentage of smokers are significantly positive for all cancers when accounting for "smoking" only, while this effect is mitigated for colorectal and larynx cancer after introducing "age" and "unemployed" sequentially. "Age" always has a positive association with incidence rates for all cancers after controlling for "smoking" and even after accounting for "unemployed".

Finally, the percentage of unemployed residents is only significantly associated with elevated incidence of colorectal cancer controlling for the other two covariates. We also find that the spatial autocorrelation ρ_i corresponding to the latent factor $\boldsymbol{f}_i$ varies considerably by cancer. Larger estimates of ρ_i imply smoother maps and, consequently, fewer difference boundaries.



(a) Smoking

(b) Smoking, Age



(c) Smoking, Age, Unemployed

Figure 3.10: Difference boundaries (highlighted in red) detected by MDAGAR after accounting for (a) smoking, (b) smoking and age, and (c) smoking, age and unemployed for four cancers individually when $\delta = 0.025$. The values in brackets are the number of difference boundaries detected.

Compared to difference boundaries for SIR in Figure 3.7a without any covariates, we tend to find lower numbers of boundaries detected with covariates included. This, too, is

Table 3.5: Posterior means (95% credible intervals) for coefficients and autocorrelation parameters estimated by adding covariates (smoking, age, unemployed) sequentially

Parameters	Lung	Esophageal	Larynx	Colorectal
Intercept	0.220 (0.177, 0.262)	0.195 (0.141, 0.254)	0.201 (0.127, 0.277)	0.197 (0.132, 0.251)
ρ_i	0.548 (0.322, 0.779)	0.965 (0.895, 0.993)	0.527 (0.051, 0.961)	0.282 (0.040, 0.695)
Intercept	0.004 (-0.022, 0.032)	-0.020 (-0.118, 0.036)	-0.010 (-0.144, 0.102)	0.084 (0.043, 0.124)
Smoking	0.022 (0.020, 0.025)	0.023 (0.017, 0.029)	0.027 (0.018, 0.039)	0.011 (0.006, 0.015)
ρ_i	0.307 (0.181, 0.439)	0.914 (0.751, 0.988)	0.513 (0.348, 0.689)	0.301 (0.011, 0.708)
Intercept	-0.064 (-0.080, -0.049)	-0.260 (-0.317, -0.214)	-0.239 (-0.335, -0.157)	0.013 (-0.015, 0.049)
Smoking	0.009 (0.006, 0.013)	0.010 (0.001, 0.021)	0.028 (0.016, 0.038)	0.003 (-0.003, 0.010)
Age	0.050 (0.046, 0.055)	0.065 (0.056, 0.074)	0.043 (0.033, 0.054)	0.042 (0.033, 0.048)
ρ_i	0.260 (0.073, 0.500)	0.589 (0.129, 0.846)	0.564 (0.025, 0.974)	0.721 (0.200, 0.960)
Intercept	-0.080 (-0.100, -0.068)	-0.28 (-0.351, -0.200)	-0.181 (-0.269, -0.101)	-0.031 (-0.047, -0.011)
Smoking	0.018 (0.005, 0.030)	0.018 (0.003, 0.033)	0.021 (-0.001, 0.040)	0.003 (-0.004, 0.009)
Age	0.048 (0.042, 0.055)	0.066 (0.056, 0.079)	0.044 (0.031, 0.055)	0.041 (0.033, 0.048)
Unemployed	-0.003 (-0.013, 0.005)	0.001 (-0.021, 0.028)	0.013 (-0.016, 0.034)	0.012 (0.005, 0.019)
ρ_i	0.302 (0.081, 0.563)	0.799 (0.481, 0.978)	0.804 (0.263, 0.987)	0.387(0.190, 0.613)

not surprising as the covariates can absorb the differences between neighboring counties and mitigate the residual effects. However, the dependencies among the cancers, the regions and the covariates is complicated and one does not always see a clear pattern. The case for "smoking" is pertinent. Figure 3.10a presents boundaries after accounting for "smoking". We see considerably fewer numbers of boundaries (eight fewer) for larynx but twelve more boundaries for esophageal cancer. The reduction in boundaries in spatial random effects can be attributed to the significant differences between smoking rates in those neighboring counties, i.e. the difference of SIR in neighboring counties is explained by the difference of smoking rates. For example, the smoking rate in Mariposa is 8.6% higher than that in Madera (19.2% vs. 10.6%). Figure 3.11 for "smoking" reflects the elimination of boundaries between the pairs of neighboring counties such as [Madera, Mariposa], [Inyo, Tulare], [Inyo, Fresno] for Larynx cancer and [Nevada, Placer] for esophageal cancer. At the same time some new boundaries appear after accounting for "smoking" including [Del Norte, Humboldt] for lung cancer, [Tulare, Fresno], [Tulare, Kings], [Tulare, Kern] and [San Bernardino, Riverside] for esophageal cancer, and [San Bernardino, Riverside] for colorectal cancer, to offset the dif-

ference of smoking rates in those pairs of counties. It implies the opposite boundary effect of other latent factors against smoking rates in those neighboring counties. Figure 3.10b reveals a considerable decrease in difference boundaries for esophageal and colorectal cancers, and to a slightly lesser extent for larynx as well, after accounting for "age". The spatial pattern for "age" in neighboring counties (see Figure 3.11) explains most boundaries for these cancers. Finally, Figure 3.10c depicts that the difference boundaries for esophageal and larynx cancers are explained by "unemployment" in neighboring counties (referring to Figure 3.11), and the number of boundaries detected for lung cancer also decrease substantially. Further discussions about cross-cancer difference boundaries are supplied in Appendix 3.7.3 of the supplementary materials.

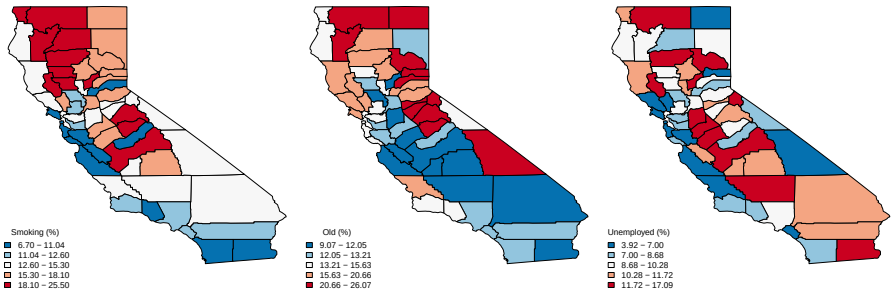


Figure 3.11: Maps of county-level covariates: adult cigarette smoking rates (left), percentage of residents older than 65 years old (middle) and unemployed residents (right).

3.6 Summary

The "MARDP" detects spatial difference boundaries for multiple correlated diseases that allows us to formulate the problem of areal boundary detection, or "areal wombling", as a Bayesian multiple testing problem for spatial random effects. Crucially, the MARDP imposes discrete probability laws on the spatial random effects and we are able to obtain fully model-based estimates of the posterior probabilities for equality of the random effects. This, in turn, allows us to use a Bayesian FDR rule to detect the boundaries.

Our data analysis on four cancers in California from the SEER database reveals that the difference boundaries vary by cancer type under the same FDR threshold. Larynx exhibits a smoother SIR map with fewer difference boundaries while more are detected for lung cancer. Covariates also impact difference boundaries for residual spatial random effects for each cancer as accounting for differences in risk factors among neighboring counties can mitigate the differences in spatial random effects. These are clearly observed for esophageal, larynx and colorectal cancers, while difference boundaries for lung cancer remain pronounced even after accounting for risk factors.

3.7 Appendix

3.7.1 Algorithm for MCMC updates

The Algorithm 1 is referenced for model implementation in Section 3.2.1.

Algorithm 1: Obtaining posterior inference of $\{\boldsymbol{\beta}_i, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{V}, \boldsymbol{\sigma}, \tau_s, \boldsymbol{\rho}, \boldsymbol{A}\}$ based on MARDP joint model

---

1. update $\boldsymbol{\beta}_i|\boldsymbol{y}_i, \boldsymbol{w}_i, \tau_i$, $\tau_i = 1/\sigma_i^2$

$$p(\boldsymbol{\beta}_i|\boldsymbol{y}_i, \boldsymbol{w}_i, \tau_i) = N\left(\left(\tau_i \boldsymbol{X}_i^\top \boldsymbol{X}_i + 1/\sigma_\beta^2 \boldsymbol{I}_{p_i}\right)^{-1} \tau_i \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \boldsymbol{w}_i), \left(\tau_i \boldsymbol{X}_i^\top \boldsymbol{X}_i + 1/\sigma_\beta^2 \boldsymbol{I}_{p_i}\right)^{-1}\right)$$

   where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i,n})^\top$ and $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n})^\top$.

2. update $\theta_k|\boldsymbol{\beta}, \boldsymbol{\tau}, \tau_s$, $k = 1, \ldots, K$, $\boldsymbol{\tau} = \{\tau_1, \tau_2, \ldots, \tau_q\}$

$$p(\theta_j|\boldsymbol{\beta}, \boldsymbol{\tau}, \tau_s) = N\left(\frac{\sum_{i=1}^q \tau_i \sum_{j:u_{ij}=k}\left(y_{ij} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_i\right)}{\sum_{i=1}^q \tau_i \sum_{j=1}^n I(u_{ij}=k) + \tau_s}, \frac{1}{\sum_{i=1}^q \tau_i \sum_{j=1}^n I(u_{ij}=k) + \tau_s}\right)$$

3. update $\gamma_{ij}|\boldsymbol{\beta}, \boldsymbol{\theta}, \tau_i, \boldsymbol{A}, \boldsymbol{\rho}$

   (a) Sample candidate γ_{ij}^* from $N(\gamma_{ij}, s_1^2)$

   (b) Compute the corresponding candidate u_{ij}^* through $u_{ij} = \sum_{k=1}^K kI\left(\sum_{t=1}^{k-1} p_t < F^{(i,j)}(\gamma_{ij}) < \sum_{t=1}^k p_t\right)$

(c) Accept γ_{ij}^* with probability

$$min\left\{1, \frac{\exp\left(-\frac{1}{2}\boldsymbol{\gamma}^{*T}\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma}^*\right) exp\left(-\frac{\tau_i}{2}\left(y_{ij}-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i-\theta_{u_{ij}^*}\right)^2\right)}{\exp\left(-\frac{1}{2}\boldsymbol{\gamma}^{T}\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma}\right)\exp\left(-\frac{\tau_i}{2}\left(y_{ij}-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i-\theta_{u_{ij}}\right)^2\right)}\right\}$$

4. update $V_k|\boldsymbol{\beta},\boldsymbol{\theta},\tau_i,\boldsymbol{\gamma}$

   (a) Sample candidate V_k^* from $N(V_k, s_2^2)$

   (b) Compute the corresponding candidate $\boldsymbol{p}^*$ and $\boldsymbol{u}^*$, where $\boldsymbol{p}=\{p_1,\ldots,p_K\}$ and $\boldsymbol{u}=\{u_1,\ldots,u_N\}$

   (c) Accept V_k^* with probability

$$min\left\{1, \frac{(1-V_k^*)^{\alpha-1}\prod_{i=1}^q\prod_{j=1}^n\exp\left(-\frac{\tau_i}{2}\left(y_{ij}-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i-\theta_{u_{ij}^*}\right)^2\right)}{(1-V_k)^{\alpha-1}\prod_{i=1}^q\prod_{j=1}^n\exp\left(-\frac{\tau_i}{2}\left(y_{ij}-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i-\theta_{u_{ij}}\right)^2\right)}\right\}$$

5. update $\tau_i|\boldsymbol{\beta},\boldsymbol{\theta}$

$$p(\tau_i|\boldsymbol{\beta},\boldsymbol{\theta})=\Gamma\left(\frac{n}{2}+a_e, \frac{1}{2}\sum_{j=1}^n\left(y_{ij}-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_i-\theta_{u_{ij}}\right)^2+b_e\right)$$

6. update $\tau_s|\boldsymbol{\theta}$

$$p(\tau_s|\boldsymbol{\theta})=\Gamma\left(\frac{K}{2}+a_s, \frac{1}{2}\sum_{k=1}^K\theta_k^2+b_s\right)$$

7. update $\boldsymbol{\rho}|\boldsymbol{\gamma}$

   (a) Let $\boldsymbol{\eta}=\text{logit}(\boldsymbol{\rho})$ and sample the candidate η_i^* from $N(\eta_i, s_3^2)$, then $\rho_i^*=\frac{exp(\eta_i^*)}{1+exp(\eta_i^*)}$

   (b) Accept $\boldsymbol{\rho}^*$ with probability

$$min\left\{1, \frac{|\boldsymbol{\Sigma}_\gamma^*|^{-\frac{N}{2}}\exp\left(-\frac{1}{2}\boldsymbol{\gamma}^T\boldsymbol{\Sigma}_\gamma^{*-1}\boldsymbol{\gamma}\right)\prod_{i=1}^q\rho_i^*(1-\rho_i^*)}{|\boldsymbol{\Sigma}_\gamma|^{-\frac{N}{2}}\exp\left(-\frac{1}{2}\boldsymbol{\gamma}^T\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma}\right)\prod_{i=1}^q\rho_i(1-\rho_i)}\right\}$$

8. update $\boldsymbol{A}|\boldsymbol{\gamma}$

   (a) Let $z_{ii}=log(a_{ii})$ and sample candidates z_{ii}^* from $N(z_{ii}, s_4^2)$

   (b) For off-diagonal elements $a_{ih}, i\neq h$, a_{ij}^* are sampled from $N(a_{ih}, s_5^2)$

(c) Accept $\boldsymbol{A}^*$ with probability

$$min\left\{1, \frac{|\boldsymbol{\Sigma}_\gamma^*|^{-\frac{N}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_\gamma^{*-1}\boldsymbol{\gamma}\right) p(\boldsymbol{A}^*) \prod_{i=1}^q a_{ii}^*}{|\boldsymbol{\Sigma}_\gamma|^{-\frac{N}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\gamma}\right) p(\boldsymbol{A}) \prod_{i=1}^q a_{ii}}\right\}$$

3.7.2 Evaluation of parameter estimates in simulation study

For the simulation study in Section 3.4, we evaluated parameter estimates from MAGAR, MCAR, DAGAR$_{ind}$ and CAR$_{ind}$ models. Table 3.6 shows the coverage probabilities (CP) defined as the coverage rates of the 95% credible intervals for each parameter over 30 datasets. All the models appear to provide effective coverages between $90\%-100\%$ for slope parameters $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12})^\top$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22})^\top$, and 100% coverage for τ_s. In terms of the variance parameters for random noise, σ_1^2 and σ_2^2, MDAGAR and MCAR offer comparable coverages at around 85%, while the two independent-disease models present much lower coverage probabilities as they fail to acquire dependent spatial structures for random effects.

The 95% credible intervals for the spatial autocorrelation parameters ρ_1 and ρ_2 estimated from CAR-based models are wide (nearly covering the entire interval $(0, 1)$). Therefore, we computed the mean squared errors (MSE) (measuring the error between estimated values and the true values) over 30 datasets instead. Table 3.7 shows estimated MSEs of ρ_1 and ρ_2 from each model. Recall that the true values $\rho_1 = 0.2$ and $\rho_2 = 0.8$. Unsurprisingly, MDAGAR delivers better inferential performance for ρ_1, while MCAR is superior for ρ_2. This finding is consistent with the finding that (univariate) DAGAR delivers better estimates of the autocorrelation parameters when ρ is not too high [DBH19].

Table 3.6: Coverage probability (%) of parameters estimated from MAGAR, MCAR, DAGAR_{ind} and CAR_{insd}

	β_{11}	β_{12}	β_{21}	β_{22}	σ_1^2	σ_2^2	τ_s
MDAGAR	100	96.7	100	93.3	83.3	86.7	100
MCAR	96.7	96.7	100	90	90	83.3	100
DAGAR_{ind}	100	96.7	100	96.7	26.7	26.7	100
CAR_{ind}	100	96.7	100	96.7	56.7	66.7	100

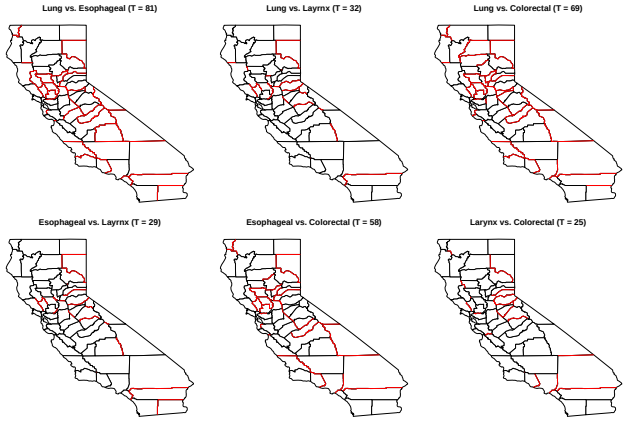
Table 3.7: Estimated MSEs of autocorrelation parameters ρ_1 and ρ_2 estimated from MAGAR, MCAR, DAGAR_{ind} and CAR_{ind}.

Methods	MSE_{ρ_1}	MSE_{ρ_2}
MDAGAR	0.034	0.173
MCAR	0.143	0.082
DAGAR_{ind}	0.590	0.028
CAR_{ind}	0.020	0.201

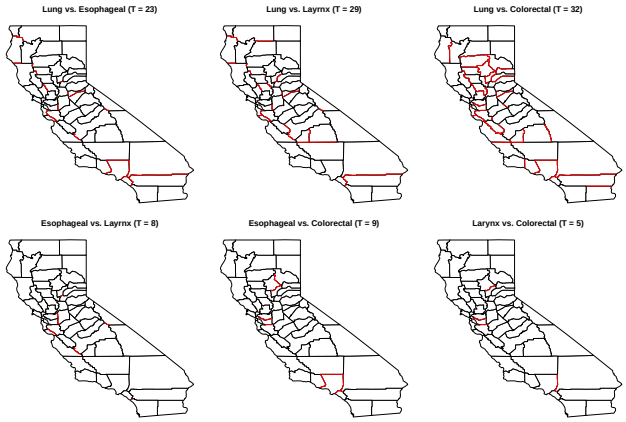
3.7.3 Impact of covariates on mutual cross-cancer difference boundaries

Figure 3.12 presents a map of California with the names and boundaries of each county. Accounting for covariates also affects the detection of mutual cross-cancer difference boundaries for each pair of cancers. Figure 3.13 shows mutual cross-cancer difference boundaries detected for each pair of cancers after accounting for only "smoking" in 3.13a, accounting for "smoking" and "age" in 3.13b and acounting for all three covariates ("smoking", "age" and "unemployed") in 3.13c when $\delta = 0.025$. Accounting only for "smoking" does not alter the cross-cancer difference boundaries as much. Most of the mutual cross-cancer difference boundaries are explained by the spatial pattern of "age" in neighboring counties, especially across esophageal, larynx and colorectal cancers where only very few boundaries between pairwise cancers are evinced from the residual spatial random effects. This is consistent with our findings from the individual cancer analysis as discussed in Section 3.5.2. Finally, accounting for "unemployment" eliminates difference boundaries further across lung, larynx and colorectal cancers, but increases the number of boundaries detected between esophageal cancer and each of the other three cancers.

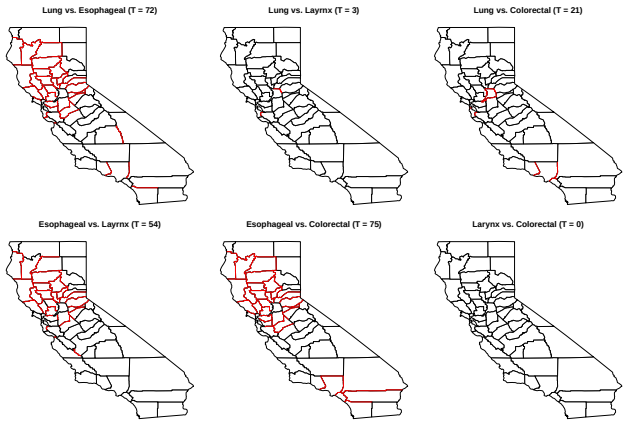
Figure 3.12: California map with county names labeled



(a) Smoking



(b) Smoking, Age



(c) Smoking, Age, Unemployed

Figure 3.13: Mutual cross-cancer difference boundaries (highlighted in red) detected by MDAGAR for each pair of cancers after accounting for (a) smoking, (b) smoking and age, and (c) smoking, age and unemployed when $\delta = 0.025$. The values in brackets are the number of difference boundaries detected.

CHAPTER 4

Discussion

In this dissertation, we introduce different hierarchical models for multivariate areal data analysis when more than one outcome (disease) observed in regional units, including multivariate disease mapping and difference boundary detection, and we illustrate with simulation studies and real data analysis using the incidence maps of California from SEER database.

In Chapter 2, based on univariate DAGAR, we developed multivariate DAGAR using conditional modelling for multivariate disease mapping, which combined with model selection approach (bridge sampling) to select the best order of MDAGAR as well as provide model weights for model averaging inference. The new method is more generalized to multiple diseases (more than two) with effective determination of the fixed order of diseases. Moreover, it provides better interpretation for spatial autocorrelations and exceeds in performance compared to GMCAR in most scenarios. However, the computation is cumbersome and hard to realize with very large number of diseases. This issue is even outstanding when it is embedded in other models like the nonparametric hierarchical models in Chapter 3. Instead, we developed another multivariate DAGAR using joint modelling without the specification of ordering for diseases.

Then in Chapter 3, we developed a multivariate nonparametric Bayesian framework "MARDP" based on an existing univariate ARDP which permits the estimation of probability that an edge being a difference boundary, and used a joint MDAGAR to construct spatial components by incorporating associations across diseases as well as space. This multivariate approach detects spatial difference boundaries not only for each disease individually but also

across associated diseases with competitive detection rates. By controlling the FDR for multiple comparison problems, it also exhibits the impact of covariates in difference boundary detection.

In the future, the direction of research will extend to the exploration of other models for multivariate difference boundary detection. For example, we can develop another Bayesian nonparametric model "MARSB" based on the areally referenced spatial sticking-breaking prior (ARSB) model [LBH15]. The comparison of different multivariate models for difference boundary detection will present a broader view on model fitting and detection efficiency.

REFERENCES

[ABS10] Jamal Akhtar, Rakesh Bhargava, Mohammad Shameem, Saurabh K Singh, Ummul Baneen, Nafees Ahmad Khan, Jassem Hassan, and Prakhar Sharma. "Second Primary Lung Cancer with Glottic Laryngeal Cancer as Index Tumor–A Case Report." *Case reports in oncology*, **3**(1):35–39, 2010.

[AMA18] Kriti Agrawal, Ronald J Markert, and Sangeeta Agrawal. "Risk factors for adenocarcinoma and squamous cell carcinoma of the esophagus and lung." *Hypertension*, **61**(46):0–09, 2018.

[Ban16] Sudipto Banerjee. "Spatial data analysis." *Annual review of public health*, **37**:47–60, 2016.

[Ban20] Sudipto Banerjee. "Modeling massive spatial datasets using a conjugate Bayesian linear modeling framework." *Spatial Statistics*, **37**:100417, 2020.

[BCG14] S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2014.

[BCL12] Sudipto Banerjee, Bradley P Carlin, Pei Li, and Alexander M McBean. "Bayesian areal wombling using false discovery rates." *Statistics and its Interface*, **5**(2):149–158, 2012.

[Bes74] Julian Besag. "Spatial interaction and the statistical analysis of lattice systems." *Journal of the Royal Statistical Society: Series B (Methodological)*, **36**(2):192–225, 1974.

[BG06] Sudipto Banerjee and Alan E Gelfand. "Bayesian Wombling: Curvilinear Gradient Assessment under Spatial Process Models." *Journal of the American Statistical Association*, **101**(476):1487–1501, 2006.

[BH95] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal statistical society: series B (Methodological)*, **57**(1):289–300, 1995.

[BHW15a] Jonathan R. Bradley, Scott H. Holan, and Christopher K. Wikle. "Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics." *Ann. Appl. Stat.*, **9**(4):1761–1791, 12 2015.

[BHW15b] Jonathan R Bradley, Scott H Holan, Christopher K Wikle, et al. "Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics." *The Annals of Applied Statistics*, **9**(4):1761–1791, 2015.

[BHW18] Jonathan R Bradley, Scott H Holan, and Christopher K Wikle. "Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data (with Discussion)." *Bayesian Analysis*, **13**(1):253–310, 2018.

[Bro64] D Brook. "On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems." *Biometrika*, **51**(3/4):481–483, 1964.

[BRT05] Nicky Best, Sylvia Richardson, and Andrew Thomson. "A comparison of Bayesian spatial models for disease mapping." *Statistical methods in medical research*, **14**(1):35–59, 2005.

[BYM91] Julian Besag, Jeremy York, and Annie Mollié. "Bayesian image restoration, with two applications in spatial statistics." *Annals of the institute of statistical mathematics*, **43**(1):1–20, 1991.

[Cal18a] California Department of Public Health. "California Tobacco Control Program. California Tobacco Facts and Figures.", 2018.

[Cal18b] California Department of Public Health, California Tobacco Control Program. "California Tobacco Facts and Figures 2018." *Sacramento, CA: California Department of Public Health*, 2018.

[CB03] Bradley P Carlin, Sudipto Banerjee, et al. "Hierarchical multivariate CAR models for spatio-temporally correlated survival data." *Bayesian statistics*, **7**(7):45–63, 2003.

[CM07] Bradley P Carlin and Haijun Ma. "Bayesian multivariate areal wombling for multiple disease boundary analysis." *Bayesian analysis*, **2**(2):281–302, 2007.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley Series in Telecommunications and Signal Processing. Wiley Interscience, 1991.

[DBH19] Abhirup Datta, Sudipto Banerjee, James S. Hodges, and Leiwen Gao. "Spatial Disease Mapping Using Directed Acyclic Graph Auto-Regressive (DAGAR) Models." *Bayesian Analysis*, **14**(4):1221 – 1244, 2019.

[DDB08] Ulysses Diva, Dipak K Dey, and Sudipto Banerjee. "Parametric models for spatially correlated survival data for individuals with multiple cancers." *Statistics in medicine*, **27**(12):2127–2144, 2008.

[DZZ06] Michael J Daniels, Zhigang Zhou, and Hui Zou. "Conditionally Specified Space-Time Models for Multivariate Processes." *Journal of Computational and Graphical Statistics*, **15**(1):157–177, 2006.

[GB19] Heli Gao and Jonathan R Bradley. "Bayesian Analysis of Areal Data with Unknown Adjacencies Using the Stochastic Edge Mixed Effects Model." *Spatial Statistics*, **31**:100357, 2019.

[GBD20] Leiwen Gao, Sudipto Banerjee, and Abhirup Datta. "Spatial Modeling for Correlated Cancers Using Bivariate Directed Graphs." *Annals of Cancer Epidemiology*, **4**(0), 2020.

[GD94] Alan E Gelfand and Dipak K Dey. "Bayesian model choice: asymptotics and exact calculations." *Journal of the Royal Statistical Society: Series B (Methodological)*, **56**(3):501–514, 1994.

[GDB21] Leiwen Gao, Abhirup Datta, and Sudipto Banerjee. "Hierarchical Multivariate Directed Acyclic Graph Auto-Regressive (MDAGAR) Models for Spatial Diseases Mapping." *arXiv preprint arXiv:2102.02911*, 2021.

[GG98] Alan E Gelfand and Sujit K Ghosh. "Model choice: a minimum posterior predictive loss approach." *Biometrika*, **85**(1):1–11, 1998.

[GHV14] Andrew Gelman, Jessica Hwang, and Aki Vehtari. "Understanding predictive information criteria for Bayesian models." *Statistics and computing*, **24**(6):997–1016, 2014.

[GL06] Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

[GSM17] Quentin F Gronau, Alexandra Sarafoglou, Dora Matzke, Alexander Ly, Udo Boehm, Maarten Marsman, David S Leslie, Jonathan J Forster, Eric-Jan Wagenmakers, and Helen Steingroever. "A tutorial on bridge sampling." *Journal of mathematical psychology*, **81**:80–97, 2017.

[GSW17] Quentin F Gronau, Henrik Singmann, and Eric-Jan Wagenmakers. "Bridgesampling: An R package for estimating normalizing constants." *arXiv preprint arXiv:1710.08162*, 2017.

[GV03] Alan E Gelfand and Penelope Vounatsou. "Proper multivariate conditional autoregressive models for spatial data analysis." *Biostatistics*, **4**(1):11–15, 2003.

[HBL15] Timothy Hanson, Sudipto Banerjee, Pei Li, and Alexander McBean. "Spatial Boundary Detection for Areal Counts." In *Nonparametric Bayesian Inference in Biostatistics*, pp. 377–399. Springer, 2015.

[HMR99] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. "Bayesian model averaging: a tutorial." *Statistical science*, pp. 382–401, 1999.

[HNF05] Leonhard Held, Isabel Natário, Sarah Elaine Fenton, Håvard Rue, and Nikolaus Becker. "Towards joint disease mapping." *Statistical methods in medical research*, **14**(1):61–82, 2005.

[JBC07] Xiaoping Jin, Sudipto Banerjee, and Bradley P Carlin. "Order-free coregionalized areal data models with application to multiple-disease mapping." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(5):817–838, 2007.

[JCB05] Xiaoping Jin, Bradley P Carlin, and Sudipto Banerjee. "Generalized hierarchical multivariate CAR models for areal data." *Biometrics*, **61**(4):950–961, 2005.

[JG03a] Geoffrey M Jacquez and Dunrie A Greiling. "Geographic boundaries in breast, lung and colorectal cancers in relation to exposure to air toxics in Long Island, New York." *International Journal of Health Geographics*, **2**(1):1–22, 2003.

[JG03b] Geoffrey M Jacquez and Dunrie A Greiling. "Local clustering in breast, lung and colorectal cancer in Long Island, New York." *International Journal of Health Geographics*, **2**(1):1–12, 2003.

[KB01] Leonhard Knorr-Held and Nicola G Best. "A shared component model for detecting joint and selective clustering of two diseases." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **164**(1):73–85, 2001.

[KC08] W Daniel Kissling and Gudrun Carl. "Spatial autocorrelation and the selection of simultaneous autoregressive models." *Global Ecology and Biogeography*, **17**(1):59–71, 2008.

[KMW18] Koich Kurishima, Kunihiko Miyazaki, Hiroko Watanabe, Toshihiro Shiozawa, Hiroichi Ishikawa, Hiroaki Satoh, and Nobuyuki Hizawa. "Lung cancer patients with synchronous colon cancer." *Molecular and clinical oncology*, **8**(1):137–140, 2018.

[Koc05] Tom Koch. *Cartographies of disease: maps, mapping, and medicine*. Esri Press Redlands, CA, 2005.

[KST01] Hoon Kim, Dongchu Sun, and Robert K Tsutakawa. "A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model." *Journal of the American Statistical association*, **96**(456):1506–1521, 2001.

[Law13] Andrew B Lawson. *Statistical methods in spatial epidemiology*. John Wiley & Sons, 2013.

[LBH15] Pei Li, Sudipto Banerjee, Timothy A Hanson, and Alexander M McBean. "Bayesian models for detecting difference boundaries in areal data." *Statistica Sinica*, pp. 385–402, 2015.

[LBM11] P Li, S Banerjee, and AM McBean. "Mining edge effects in areally referenced spatial data: A Bayesian model choice approach." *Geoinformatica*, **15**:435–454, 2011.

[LC05] Haolan Lu and Bradley P Carlin. "Bayesian areal wombling for geographical boundary analysis." *Geographical Analysis*, **37**(3):265–285, 2005.

[LFB17] Sara Lindström, Hilary Finucane, Brendan Bulik-Sullivan, Fredrick R Schumacher, Christopher I Amos, Rayjean J Hung, Kristin Rand, Stephen B Gruber, David Conti, Jennifer B Permuth, et al. "Quantifying the genetic correlation between multiple cancer types." *Cancer Epidemiology and Prevention Biomarkers*, **26**(9):1427–1435, 2017.

[Mac18] Ying C. MacNab. "Some recent work on multivariate Gaussian Markov random fields (with discussion)." *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, **27**(3):497–541, September 2018.

[Mar88] KV Mardia. "Multi-dimensional multivariate Gaussian Markov random fields with application to image processing." *Journal of Multivariate Analysis*, **24**(2):265–284, 1988.

[Mar13] Miguel A Martinez-Beneito. "A general modelling framework for multivariate disease mapping." *Biometrika*, **100**(3):539–553, 2013.

[MBB17] Miguel A Martinez-Beneito, Paloma Botella-Rocamora, and Sudipto Banerjee. "Towards a multidimensional approach to Bayesian disease mapping." *Bayesian analysis*, **12**(1):239, 2017.

[MCB10] Haijun Ma, Bradley P Carlin, and Sudipto Banerjee. "Hierarchical and joint site-edge methods for Medicare hospice service region boundary analysis." *Biometrics*, **66**(2):355–364, 2010.

[MMG14] Marc Marí-DellOlmo, Miguel A Martinez-Beneito, Mercè Gotsens, and Laia Palència. "A smoothed ANOVA model for multivariate ecological regression." *Stochastic environmental research and risk assessment*, **28**(3):695–706, 2014.

[MPR04] Peter Müller, Giovanni Parmigiani, Christian Robert, and Judith Rousseau. "Optimal sample size for multiple testing: the case of gene expression microarrays." *Journal of the American Statistical Association*, **99**(468):990–1001, 2004.

[MW96] Xiao-Li Meng and Wing Hung Wong. "Simulating ratios of normalizing constants via a simple identity: a theoretical exploration." *Statistica Sinica*, pp. 831–860, 1996.

[Nat19] National Cancer Institute. "SEER*Stat Software.", Aug 2019.

[PGV04] Marco Perone Pacifico, Christopher Genovese, Isabella Verdinelli, and Larry Wasserman. "False discovery control for random fields." *Journal of the American Statistical Association*, **99**(468):1002–1014, 2004.

[PNT14] Konstantinos Perrakis, Ioannis Ntzoufras, and Efthymios G Tsionas. "On the use of marginal posteriors in marginal likelihood estimation via importance sampling." *Computational Statistics & Data Analysis*, **77**:54–69, 2014.

[QBN21] Kai Qu, Jonathan R Bradley, and Xufeng Niu. "Boundary Detection Using a Bayesian Hierarchical Model for Multiscale Spatial Data." *Technometrics*, **63**(1):64–76, 2021.

[RC13] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[RH05] Havard Rua and Leonard Held. *Gaussian Markov Random Fields : Theory and Applications*. Monographs on statistics and applied probability. Chapman and Hall/CRC Press, Boca Raton, FL, 2005.

[SC04] Wei-Xing Shi and Shu-Qing Chen. "Frequencies of poor metabolizers of cytochrome P450 2C19 in esophagus cancer, stomach cancer, lung cancer and bladder cancer in Chinese population." *World journal of gastroenterology: WJG*, **10**(13):1961, 2004.

[Sch13] Wesley L Schaible. *Indirect estimators in US federal programs*, volume 108. Springer Science & Business Media, 2013.

[see] "Static County Attributes - SEER Datasets.".

[Set94] Jayaram Sethuraman. "A constructive definition of Dirichlet priors." *Statistica sinica*, pp. 639–650, 1994.

[TKP18] Wesley Tansey, Oluwasanmi Koyejo, Russell A Poldrack, and James G Scott. "False Discovery Rate Smoothing." *Journal of the American Statistical Association*, **113**(523):1156–1171, 2018.

[Wal04] Melanie M Wall. "A close look at the spatial structure implied by the CAR and SAR models." *Journal of statistical planning and inference*, **121**(2):311–324, 2004.

[Wat10] Sumio Watanabe. "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of Machine Learning Research*, **11**(Dec):3571–3594, 2010.

[WC10] L Waller and B Carlin. "Handbook Of Spatial Statistics.", 2010.

[WCX97] Lance A Waller, Bradley P Carlin, Hong Xia, et al. "Hierarchical spatio-temporal mapping of disease rates." *Journal of the American Statistical association*, **92**(438):607–617, 1997.

[WG04] Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*, volume 368. John Wiley & Sons, 2004.

[Wom51] William H Womble. "Differential systematics." *Science*, **114**(2961):315–322, 1951.

[ZEY05] J. Zhu, J. C. Eickhoff, and P. Yan. "Generalized Linear Latent Variable Models for Repeated Measures of Spatially Correlated Multivariate Data." *Biometrics*, **61**(3):674–683, 2005.

[ZHB09] Yufen Zhang, James S Hodges, and Sudipto Banerjee. "Smoothed ANOVA with spatial effects as a competitor to MCAR in multivariate spatial smoothing." *The annals of applied statistics*, **3**(4):1805, 2009.