PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Fast Bayesian Inference for multivariate DAGAR Models

Tesis para optar por el grado académico de Maestro en Estadística que presenta:

Renzo Jesús Moscoso Basaldúa

ASESORA:

Zaida Jesús Quiroz Cornejo

Lima, 2025

INFORME DE SIMILITUD

Yo Zaida Jesús Quiroz Cornejo docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesora de la tesis titulada *Fast Bayesian Inference for multivariate DAGAR Models*, del autor Renzo Jesús Moscoso Basaldúa, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 22%. Así lo consigna el reporte de similitud emitido por el software Turnitin el 22/08/2025.
- He revisado con detalle dicho reporte y confirmo que cada una de las coincidencias detectadas no constituyen plagio alguno.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 22 de agosto de 2025

<table>
<tr><td colspan="2">Apellidos y nombres de la asesora:
Quiroz Cornejo Zaida Jesús</td></tr>
<tr><td>DNI: 43704124</td><td rowspan="2">Firma:</td></tr>
<tr><td>ORCID: https://orcid.org/0000-0003-3821-0815</td></tr>
</table>

Acknowledgments

I would like to express my deepest gratitude to my thesis advisor, Professor Zaida Quiroz, for her invaluable guidance, patience, and support throughout this research journey. Her expertise, insightful feedback, and constant encouragement were instrumental in shaping this work.

Resumen

El modelamiento de datos de áreas en un contexto multivariado es fundamental para evaluar con precisión la relación entre las variables respuesta, así como su dependencia espacial. Asimismo, una estimación precisa de la autocorrelación espacial es útil para identificar patrones espaciales y mejorar el modelado estadístico en campos como la epidemiología, la ecología y la planificación urbana. Sin embargo, los modelos convencionales para la estimación de datos de áreas multivariados, generalmente extensiones del modelo autorregresivo condicional (CAR), tienden a sobreestimar el parámetro asociado a la autocorrelación espacial. En este contexto, esta tesis propone un modelo para datos de áreas multivariados con el objetivo de mejorar la estimación de la autocorrelación espacial en un entorno de múltiples variables respuesta. Este modelo es una extensión del modelo autorregresivo de grafo acíclico dirigido (DAGAR), el cual ha demostrado ofrecer una mejor estimación del parámetro de autocorrelación espacial en comparación con los modelos CAR. El modelo propuesto se implementa bajo un enfoque bayesiano utilizando la Aproximación Laplaciana Anidada Integrada (INLA) para mayor eficiencia computacional. Finalmente, para evaluar la contribución de esta propuesta, se ajusta el modelo a datos reales y se compara su desempeño con un enfoque alternativo.

Palabras-clave: DAGAR, Datos de áreas, Inferencia Bayesiana, INLA, Modelo espacial multivariado.

Abstract

The modeling of areal data in a multivariate setting is important to accurately assess the relationship between the response variables as well as their spatial dependence. Furthermore, a precise estimation of the spatial autocorrelation is useful to identify spatial patterns and improve statistical modeling in fields such as epidemiology, ecology and urban planning. Nevertheless, conventional models for the estimation of multivariate areal data, usually extensions of the Conditional Autoregressive (CAR) model, tend to overestimate the parameter associated to the spatial autocorrelation. In this context, we propose a model for multivariate areal data with the purpose of improving the estimation of the spatial autocorrelation in a multivariate setting. This model is an extension of the Directed Acyclic Graph Autoregressive (DAGAR) model which showed better estimation of the spatial autocorrelation parameter when compared to CAR models. The proposed model is implemented under the Bayesian approach using the Integrated Nested Laplace Approximation (INLA) for computational efficiency. Finally, to evaluate the contribution of this proposal, we fit the model to real world data and compare its performance with a competing approach.

Keywords: DAGAR, Bayesian inference, Areal data, INLA, Multivariate model.

Contents

Chapter 1

Introduction

1.1 Preliminaries

There exist plenty of cases where we are interested in modelling the association between two or more response variables in space. These models have been proved to yield better performance than univariate models by sharing information between response variables that are correlated (Tesema et al., 2023). For example, regarding epidemiology, Adeyemi et al. (2019) investigated the underlying risk factors of three children malnutrition outcomes (anemia, stunting and wasting) in provinces of Burkina Faso and Mozambique using a multivariate conditional autoregressive model. In transportation, Huang et al. (2017) proposed a multivariate spatial model to identify factors contributing to crash risks at urban intersections for pedestrians, cyclists and motor vehicles, simultaneously, and to assess the correlation of crash counts between the three transportation modes and spatial correlation between adjacent intersections. Both studies compared the proposed multivariate model with an independent univariate model for each response variable and concluded that the multivariate model yields more precise estimates than the univariate ones.

The data available regarding these variables usually come in the form of aggregated data from different areas divided by regional boundaries. This type of data, known as areal or lattice data, has been commonly modelled, in its univariate version, by areal data models (Aswi et al., 2019). In these models, the spatial information is encoded as neighborhood matrices, and they assume that the response variable at one location is explained by both the covariates and the response variable at other locations (Lichstein et al., 2002). The most frequently used areal data models are conditional autoregressive models (CAR) and simultaneous autoregressive models (SAR), (Cressie, 2015). In particular, CAR models were introduced by Besag (1974), and are commonly used as a prior distribution for spatial random

effects in Bayesian hierarchical models. The joint distribution of the random effects of a CAR model is obtained from the conditional distribution on other random effects, specifically by assuming that the conditional probability of a region only depends on its neighbors (Schmidt and Nobre, 2018). On the other hand, SAR models were introduced by Whittle (1954) and their definition assumes a regression model on the values from its neighbors areas to account for the spatial dependence (Bivand et al., 2008). Both the SAR and CAR models usually depend on an extra parameter (ρ) which can be viewed as a spatial autocorrelation parameter, nevertheless this parameter is difficult to interpret as it tends to overestimate the spatial correlation (Banerjee et al., 2014).

Recently, Datta et al. (2019) proposed a new approach, referred to as DAGAR (Directed Acyclic Graph Auto-Regressive) model, to deal with the lack of interpretability derived from the CAR and SAR models, using directed acyclic graphs (DAGs). A DAG is a type of graph that contains no cycles, that is, it is not possible to return to any node after traveling any path that comes off that node (VanderWeele and Robins, 2007), and it is used to show how the elements of a set are related to each other. In a DAG, each edge has a direction that indicates a one-way relationship between nodes. In DAGAR models, DAGs are used to represent spatial dependence between areal units. The random effects of these areas are modeled by a Gaussian distribution. The random effect of an areal unit is only affected by a set of relevant neighbors using local spanning trees. DAGAR models showed better accuracy in estimating the latent spatial random effects and improved interpretability of the spatial correlation parameter. Furthermore, DAGAR models showed better performance under weak spatial correlation compared to CAR models.

For multivariate spatial analysis some models have been proposed (Mardia, 1988; Knorr-Held, 2000; Gelfand and Vounatsou, 2003; Martinez-Beneito, 2013; Martínez-Beneito and Botella-Rocamora, 2019). Most of them extend the univariate CAR model. For instance, Martinez-Beneito (2013) proposed a flexible multivariate CAR model, which reformulates Kronecker products of previous multivariate CAR models as simple matrix products, thus it is computationally convenient. This model allows to combine spatial dependence structures and different relationships between response variables. In general they assumed a similar spatial pattern between response variables.

Bayesian inference is typically carried out by Markov Chain Monte Carlo (MCMC) methods (Gelfand and Smith, 1990), such as the Gibbs sampling, an iterative algorithm whose simulated Markov chains converges to the desired joint posterior distribution (Hoff, 2009). However, the iterative nature of MCMC makes this algorithm of slow convergence. In that

context, Rue et al. (2009) showed a more efficient approach for estimating the posterior marginal distributions through the Integrated Nested Laplace Approximation (INLA). This approach showed to reduce the computational cost in several orders of magnitude. In particular, Palmí-Perales et al. (2021) estimated the parameters of the multivariate spatial model proposed by Martinez-Beneito (2013) through INLA.

In this thesis, a new model for multivariate spatial analysis is proposed. This model extends the models presented by Martinez-Beneito (2013) by considering the DAGAR models. A Bayesian inference approach is adopted using INLA. The implementation of the model is assessed through simulation studies for different values of the associated parameters. Finally, a real application of the model is presented, e.g. to study the relationship between two or more diseases that might be dependent and have similar spatial patterns.

1.2 Goals

In this thesis, we propose to extend the models presented by Martinez-Beneito (2013) using the DAGAR approach (Datta et al., 2019), to fit two or more response variables in space and to implement the inference of these multivariate DAGAR models through INLA. In particular:

- Propose, study properties, and implement the estimation of the multivariate DAGAR model from a Bayesian approach.
- Implement Bayesian inference methods considering INLA simulation.
- Carry out simulation studies on the multivariate DAGAR model considering intensive computing on different scenarios.
- Apply the proposed model to real data and compare the fit with other competing models.

1.3 Structure of the thesis

The organization of this thesis is as follows. Chapter 2 describes some important concepts related to areal data models, such as: areal data, spatial autocorrelation, spatial processes, graphs, adjacency matrix, that will be helpful understand the next chapters. In addition, CAR and SAR models, Bayesian inference, and INLA are explained. In Chapter 3, the proposed model (i.e, Multivariate DAGAR model for areal data) is defined and it is explained

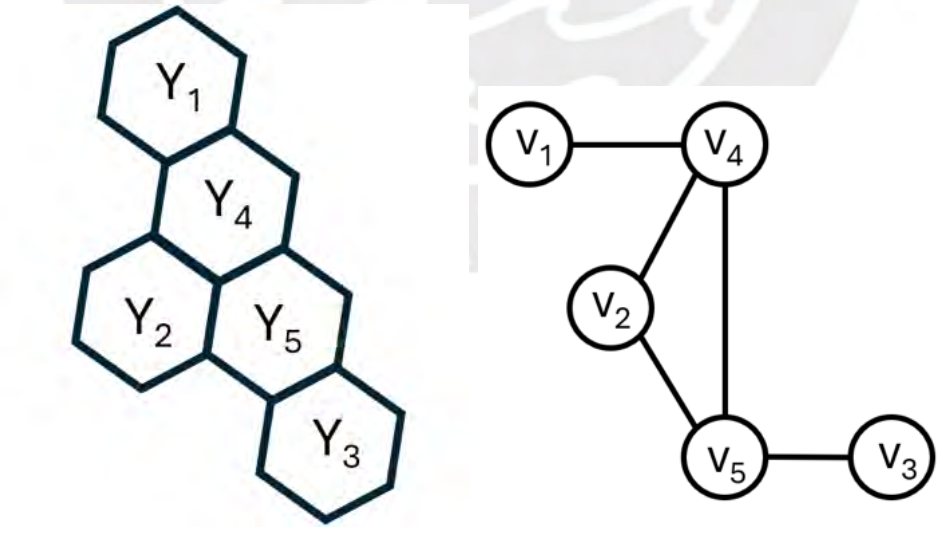
its estimation using the INLA method. In Chapter 4, a simulation study under different scenarios is presented, with the aim of assess the suitability of the model and compare its performance with the multivariate CAR model for areal data. Chapter 5 shows applications of the proposed model using real data. Finally, Chapter 6 presents the conclusions and future work.

Chapter 2

Concepts

2.1 Graphs

In order to analyze areal data, geographic domain is usually seen as an undirected graph where the areal units are represented by nodes which are connected by edges, usually defined between areas that share a border. Formally, a graph can be represented by a pair $G = \{V, E\}$, where V is a set of nodes and E is a set of edges. If G is defined as an undirected graph, an edge (v_i, v_j) joins the nodes v_i and v_j. For example, Figure 2.1a shows a region divided in five areal units represented by hexagons. Figure 2.1b shows an acyclic graph composed of five nodes $V = \{v_1, v_2, v_3, v_4, v_5\}$ and five edges $E = \{(v_1, v_4), (v_4, v_2), (v_2, v_5), (v_4, v_5), (v_5, v_3)\}$ representing the region shown in Figure 2.1a.



(a) (b)

Figure 2.1: Areal data representation. (a) Draft of a region composed of 5 areal units. (b) Corresponding graph of the region.

2.2 Adjacency matrix

An adjacency matrix ($\boldsymbol{W}$) is used to represent information about spatial proximity between n areas, thus its dimension is $n \times n$. The entries w_{ij} of $\boldsymbol{W}$, corresponds to a measure of proximity between the areas i and j. For the remainder of this chapter, we are defining w_{ij} as an indicator function as follows:

$$w_{ij} = \begin{cases} 1; & \text{if area } i \text{ shares a boundary with area } j, \text{ for } i \neq j \\ 0; & \text{otherwise.} \end{cases}$$

In other words, $w_{ij} = 1$ if there is an edge between nodes v_i and v_j in the graph of the corresponding region. As an example, the adjacency matrix, according to this definition, of the graph in Figure 2.1b is

$$\boldsymbol{W} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

2.3 Spatial autocorrelation

Spatial autocorrelation measures the degree of correlation of a random variable through the space. A positive spatial autocorrelation indicates similar values of the variable in neighbor areas, while a negative spatial autocorrelation indicates opposite or dissimilar values in areas that are closer together. There are many statistics used as indicators of spatial autocorrelation. The most commonly used are Morans' I and Geary's C, which are indicators of global and local spatial autocorrelation, respectively. Such statistics assume homogeneity over the whole area under study, i.e., the same degree of spatial autocorrelation. A brief description of these statistics is presented:

Moran's I

This statistic was proposed by Cliff and Ord (1981) based on the work of Moran (1950). It is defined as:

$$I = \frac{N}{S_0} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2},$$

where N is number of areal units, x is observed value of the random variable under study, $\bar{x}$ is the mean of x, w_{ij} are the elements of an adjacency matrix and S_0 is the sum of all w_{ij}. It

ranges from -1 to 1, where values close to 1 indicate strong positive spatial autocorrelation, values close to -1 indicate strong negative spatial autocorrelation, and values around 0 suggest no spatial autocorrelation.

Geary's C

This statistic was introduced by Geary (1954). It is defined as:

$$C = \frac{(N-1)\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - x_j)^2}{2S_0 \sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

A value approaching 0 indicates strong positive spatial autocorrelation, and a value around 1 indicates no spatial autocorrelation.

2.4 Areal data models

Areal data models are defined as generalized linear mixed models (GLMM), where the response variable in area i, Y_i, is modeled using any suitable distribution from the exponential family, $Y_i \sim FE(\mu_i, \gamma)$, where μ_i is the mean of Y_i, and spatial information is usually incorporated in the form of spatial random effects. In a general way, the mean can be associated to the linear predictor η_i as follows:

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \theta_i; \quad i = 1, 2, \ldots, n$$

where $g(.)$ is any suitable link function, $\boldsymbol{x}_i$ is the vector of covariates of the area i, $\boldsymbol{\beta}$ is the vector of fixed coefficients, θ_i is the spatial random effect of the area i. The vector of spatial random effects $\boldsymbol{\theta}$, jointly follow a multivariate Gaussian distribution as follows:

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^\top \sim N(\mathbf{0}, \boldsymbol{Q}^{-1}), \tag{2.1}$$

where $\boldsymbol{Q}$ is the precision matrix (inverse of covariance matrix) of $\boldsymbol{\theta}$.

Three models for areal data following the structure of equation 2.1 are presented in the next subsections, where the only difference resides in the form of the precision matrix $\boldsymbol{Q}$.

2.4.1 Conditional Autoregressive model (CAR)

The CAR model was introduced by Besag (1974). This model assumes that θ_i is only affected by its neighbors, then the conditional density distribution of θ_i given the rest of

spatial random effects is defined as follows:

$$f(\theta_i|\theta_j, j \neq i) = f(\theta_i|\theta_j, j \in N(i)),$$

where $N(i)$ represents the set of neighbors of area i. For example, in Figure 2.1, the full conditional distribution of θ_5 depends on $(\theta_1, \theta_2, \theta_3, \theta_4)^T$, but since θ_i is only affected by its neighbors, then θ_5 depends only on $(\theta_2, \theta_3, \theta_4)^\top$.

The full conditional distribution of a random vector can be define by its joint distribution but certain conditions are needed for the converse to be true. In particular, let define the vector $\boldsymbol{\theta_{-i}} = (\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)^\top$, then it is assumed that the full conditional distribution $\theta_i|\boldsymbol{\theta_{-i}}$ follows a normal distribution as follows:

$$\theta_i|\boldsymbol{\theta_{-i}} \sim N(\sum_{j \neq i} b_{ij}\theta_j, \sigma_i^2), \tag{2.2}$$

where b_{ij} is a constant and σ_i^2 is a marginal variance. Then the Brook's lemma allows to calculate the joint distribution in terms of the full conditional distributions, up to a normalizing constant (Besag, 1974). Specifically, using the Brook's lemma:

$$\frac{f(\boldsymbol{\theta})}{f(\boldsymbol{\theta_0})} = \prod_{i=1}^{n} \frac{f(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{(i+1,0)}, \ldots, \theta_{(n,0)})}{f(\theta_{(i,0)}|\theta_1, \ldots, \theta_{i-1}, \theta_{(i+1,0)}, \ldots, \theta_{(n,0)})}, \tag{2.3}$$

where $\boldsymbol{\theta_0} = (\theta_{(1,0)}, \ldots, \theta_{(n,0)})$ is any fixed point in the support of $\boldsymbol{\theta}$.

Applying Brook's lemma given in equation (2.3) to the full conditional distributions proposed in equation (2.2), we obtain:

$$f(\boldsymbol{\theta}) \propto \exp\{-\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{\theta}\}, \tag{2.4}$$

where $\boldsymbol{D} = diag(\sigma_1^2, \ldots, \sigma_n^2)$ and $\boldsymbol{B} = (b_{ij})_{n \times n}$.

Let $\boldsymbol{Q} = \boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{B})$. The the expression given in equation (2.4) follows a normal distribution with zero mean and precision matrix $\boldsymbol{Q}$, if $\boldsymbol{Q}$ is symmetric and invertible. The matrix $\boldsymbol{Q}$ becomes symmetric by making $b_{ij} = \frac{w_{ij}}{w_{i+}}$ and $\sigma_i^2 = \frac{\sigma^2}{w_{i+}}$, where w_{ij} is an element of the adjacency matrix $\boldsymbol{W}$ and w_{i+} is the number of neighbors of area i. Then, $\boldsymbol{Q}$ can be expressed as,

$$\boldsymbol{Q} = \frac{1}{\sigma^2}(\boldsymbol{D_w} - \boldsymbol{W}),$$

where $\boldsymbol{D_w} = diag(w_{i+})$. This arrangement generates a singular $\boldsymbol{Q}$ matrix, which produces

an improper joint distribution for $\boldsymbol{\theta}$ called improper CAR (ICAR) model. To obtain an invertible matrix $\boldsymbol{Q}$, the full conditional mean is defined by incorporating a new parameter, ρ, such that $E(\theta_i|\boldsymbol{\theta}_{-i}) = \rho\frac{\sum_{j\neq i} w_{ij}\theta_j}{w_{i+}}$. Therefore, the full conditional distribution of θ_i can be expressed as follows:

$$\theta_i|\boldsymbol{\theta_{-i}} \sim N(\rho\frac{\sum_{j\neq i} w_{ij}\theta_j}{w_{i+}}, \frac{\sigma^2}{w_{i+}}).$$

Finally, $\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{Q})$ where $\boldsymbol{Q}$ takes the following form:

$$\boldsymbol{Q} = \frac{1}{\sigma^2}(\boldsymbol{D_w} - \rho\boldsymbol{W}),$$

where $|\rho| < 1$ so that $\boldsymbol{Q}$ is positive definite and invertible.

It can be observed that the conditional mean is a weighted average of the neighbors of area i and the conditional variance is inversely proportional to its number of neighbors. Despite the fact that ρ is usually referred to as "spatial autocorrelation parameter", it has been shown that this parameter tends to overestimate the actual spatial autocorrelation of the data, specially when the true spatial autocorrelation is small (Wall, 2004).

2.4.2 Simultaneous Autoregressive model (SAR)

The SAR model was introduced by Whittle (1954). Let θ_i be a linear combination of θ_j for $j \neq i$ as follows:

$$\theta_i = \rho\sum_{j\neq i} b_{ij}\theta_j + \varepsilon_i; \quad i = 1, 2, \ldots, n,$$

where $\varepsilon_i \overset{ind}{\sim} N(0, \sigma_i^2)$ and $b_{ij} = \frac{w_{ij}}{w_{i+}}$ as defined in the CAR model. It is assumed that ε_i is independent of θ_i. Thus the SAR model can also be expressed in matrix form as:

$$\boldsymbol{\theta} = \rho\boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \overset{ind}{\sim} N(\mathbf{0}, diag(\sigma_i^2))$ and $\boldsymbol{B}$ has elements b_{ij}. Then, $\boldsymbol{\theta}$ can be defined as:

$$\boldsymbol{\theta} = (\boldsymbol{I} - \rho\boldsymbol{B})^{-1}\boldsymbol{\varepsilon}.$$

By properties of the multivariate normal distribution:

$$\boldsymbol{\theta} \sim N(\mathbf{0}, (\boldsymbol{I} - \rho\boldsymbol{B})^{-1}\boldsymbol{F}((\boldsymbol{I} - \rho\boldsymbol{B})^{-1})^T),$$

where $\boldsymbol{F} = diag(\sigma_i^2)$. Finally, the precision matrix $\boldsymbol{Q}$ of the SAR model is expressed as follows:

$$\boldsymbol{Q} = (\boldsymbol{I} - \rho\boldsymbol{B})^T \boldsymbol{F}^{-1}(\boldsymbol{I} - \rho\boldsymbol{B}).$$

2.4.3 Directed Acyclic Graph Autoregressive model (DAGAR)

This model was proposed by Datta et al. (2019). They proposed a model based on a directed acyclic graph representation of the spatial dependence for the response variable. They achieved this by constructing a multivariate Gaussian distribution starting from a sparse Cholesky factor.

Given the vector of spatial random effects $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^\top$, its joint probability density function can be expressed by the product rule as

$$f(\boldsymbol{\theta}) = f(\theta_1)f(\theta_2|\theta_1)f(\theta_3|\theta_2, \theta_1)...f(\theta_n|\theta_{n-1}, ..., \theta_1).$$

Then, let assume that θ_i is a linear combination of θ_j for $j < i$ so that:

$$\begin{aligned}
\theta_1 &= \varepsilon_1 \\
\theta_2 &= b_{21}\theta_1 + \varepsilon_2 \\
&\vdots \\
\theta_n &= b_{n1}\theta_1 + b_{n2}\theta_2 + \cdots + b_{n,n-1}\theta_{n-1} + \varepsilon_n,
\end{aligned}$$

which can be expressed as

$$\theta_i = \sum_{j<i} b_{ij}\theta_j + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, \sigma_i^2)$ and b_{ij} is a constant. In matrix form, $\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$,where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{F})$, $\boldsymbol{B} = (b_{ij})_{n\times n}$ is a strictly lower triangular matrix, and $\boldsymbol{F} = diag(\sigma_i^2)_{n\times n}$. Then, $\boldsymbol{\theta} = (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\varepsilon}$ and $\boldsymbol{\theta} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{F}((\boldsymbol{I} - \boldsymbol{B})^{-1})^\top$. It is desirable that matrix $\boldsymbol{B}$ is sparse for computational efficiency. This is achieved by redefining the restriction imposed for the linear combination by adding the condition that j must be a neighbor of i, which can be expressed in the following way:

$$\theta_i = \sum_{j\in N^*(i)} b_{ij}\theta_j + \varepsilon_i, \qquad N^*(i) = \{j|j < i, j \in N(i)\}. \tag{2.5}$$

where $N(i)$ represents the set of neighbors of area i. In order to define $\boldsymbol{B}$ and $\boldsymbol{F}$, let G be the graph composed of all the nodes and edges using the restriction $N^*(i)$. Then G_i is

the subgraph of G composed of nodes $\{i\} \cup N^*(i)$ and their edges, and T_i is an embedded spanning tree of G_i defined as:

$$T_i = (\{i\} \cup N^*(i), \{i \sim j | j \in N^*(i)\}),$$

where $i \sim j$ indicates that area i is neighbor of area j , i.e., T_i contains only the edges between node i and nodes in $N^*(i)$.

Now, the conditional distribution of $\theta_i|\boldsymbol{\theta_{N^*(i)}}$ is designed using an autoregressive model of lag 1 (AR(1) model) on T_i with parameter ρ. The main reason is that according to Basseville et al. (1992), there is a valid covariance function on a tree graph, thus for $|\rho| < 1$, we can define the matrix $\rho^D = (\rho^{d_{kl}})$ as a covariance matrix of the joint distribution of $(\theta_i, \boldsymbol{\theta_{N^*(i)}})$, where d_{kl} is the length of the shortest path in T_i between nodes k and l. The resulting autoregressive covariance matrix of $(\theta_i, \boldsymbol{\theta_{N^*(i)}})$ is:

$$\begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho^2 & \cdots & \rho^2 \\ \rho & \rho^2 & 1 & \cdots & \rho^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho & \rho^2 & \cdots & \rho^2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{v_i^T} \\ \boldsymbol{v_i} & \boldsymbol{\Sigma_i} \end{pmatrix}. \tag{2.6}$$

Since $\boldsymbol{\theta}$ follows a multivariate normal distribution, any conditional distribution is also normal, that is

$$\theta_i|\boldsymbol{\theta_{N^*(i)}} \sim N(\mu_{i|N^*(i)}, \sigma^2_{i|N^*(i)}),$$

and from equation (2.6) the conditional mean and variance of $\theta_i|\boldsymbol{\theta_{N^*(i)}}$ are given by

$$\mu_{i|N^*(i)} = \boldsymbol{v_i^T \Sigma_i^{-1} \theta_{N^*(i)}}, \tag{2.7}$$

$$\sigma^2_{i|N^*(i)} = 1 - \boldsymbol{v_i^T \Sigma_i^{-1} v_i}.$$

On the other hand, from equation (2.5) we obtain

$$\mu_{i|N^*(i)} = E(\theta_i|\boldsymbol{\theta_{N^*(i)}}) = \sum_{j \in N^*(i)} b_{ij}\theta_j, \tag{2.8}$$

$$\sigma^2_{i|N^*(i)} = V(\theta_i|\boldsymbol{\theta_{N^*(i)}}) = \sigma_i^2.$$

From (2.7) and (2.8), it follows:

$$b_{ij} = \frac{\rho}{1+(n_{<i}-1)\rho^2}; \quad (i = 2, \ldots, n,\ j \in N^*(i)), \tag{2.9}$$
$$\tau_i = (\sigma_i^2)^{-1} = \frac{1+(n_{<i}-1)\rho^2}{1-\rho^2}; \quad (i = 1, \ldots, n),$$

where $n_{<i}$ denotes the cardinality of $N^*(i)$. Then the distribution of $\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{\sigma^2\Sigma})$, where $\boldsymbol{\Sigma} = (\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{F}((\boldsymbol{I}-\boldsymbol{B})^{-1})^T$ and σ^2 is a marginal variance. The precision matrix (inverse of covariance matrix $\boldsymbol{\Sigma}$) of $\boldsymbol{\theta}$ can be represented as follows $\boldsymbol{Q} = \boldsymbol{L}^T\boldsymbol{F}^{-1}\boldsymbol{L}$, where $\boldsymbol{L} = (\boldsymbol{I}-\boldsymbol{B})$.

As there is no natural order for the areas, there could be $n!$ orderings. To deal with this, Datta et al. (2019) proposed to define any ordering as:

$$\pi = \{\pi(1), \pi(2), \ldots, \pi(n)\}.$$

Then, $\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{\sigma^2[Q^\star]^{-1}})$, the precision matrix $\boldsymbol{Q^\star}$ free of any ordering of $\boldsymbol{\theta}$, can be calculated by averaging the Cholesky decomposition of the precision matrix of $\boldsymbol{\theta}$ for any ordering π, that is:

$$\boldsymbol{Q^\star} = \frac{1}{n!}\sum_{\pi} \boldsymbol{P}_\pi^\top \boldsymbol{L}_\pi^\top \boldsymbol{F}_\pi^{-1} \boldsymbol{L}_\pi \boldsymbol{P}_\pi,$$

where $\boldsymbol{L}_\pi = (\boldsymbol{I}-\boldsymbol{B}_\pi)$ and $\boldsymbol{P}_\pi(\theta_1, \ldots, \theta_n)^\top = (\theta_{\pi(1)}, \ldots, \theta_{\pi(n)})^\top$ for any n-dimensional vector $\boldsymbol{\theta}$ is a permutation matrix corresponding to π. This achieves that $\boldsymbol{Q^\star}$ is order free and a function of the directed acyclic graph G. The components of $\boldsymbol{B_\pi}$ and $\boldsymbol{F_\pi}$ are

$$b_{ij} = \frac{\rho}{1+(n_{\pi_{(i)}}-1)\rho^2}; \quad \tau_i = (\sigma_i^2)^{-1} = \frac{1+(n_{\pi(i)}-1)\rho^2}{1-\rho^2}, \tag{2.10}$$

where $n_{\pi(i)}$ denotes the cardinality of $N_\pi(i)$, that is, the number of neighbors of an area i.

Simulations showed improved the estimation of the spatial autocorrelation parameter ρ of this model with respect to the CAR and SAR models, which tend to overestimate the spatial autocorrelation.

2.5 Bayesian inference

Let Y be a random variable that follows a probability distribution dependent on a vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^\top$. Given a random sample $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$, we aim to obtain an estimate of $\boldsymbol{\theta}$. Under the frequentist approach, $\boldsymbol{\theta}$ would be estimated through maximum likelihood estimation (MLE), analytically or numerically, which would give us the

value of $\boldsymbol{\theta}$ that maximizes the likelihood.

In contrast to the frequentist approach, in Bayesian inference $\boldsymbol{\theta}$ is seen as a random vector that follows a distribution, $p(\boldsymbol{\theta})$, that reflects our previous knowledge about $\boldsymbol{\theta}$. This distribution is known as prior distribution. Furthermore, this knowledge about $\boldsymbol{\theta}$ can be updated, which will result in the distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, referred to as posterior distribution. Bayes' theorem gives us the posterior distribution of $\boldsymbol{\theta}$ as a function of its prior distribution $p(\boldsymbol{\theta})$ and the data likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ as follows:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})},$$

where $p(\boldsymbol{y})$ is the probability density function of $\boldsymbol{y}$. It is usual to express the posterior distribution in terms of proportionality as:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

When analytical solutions are not feasible or when the model complexity is too high, Markov Chain Monte Carlo (MCMC) methods gives us an approximation of the posterior distribution. For example, the Gibbs sampler gives us an approximation of the posterior distribution in multiparameter models when the joint posterior distribution is difficult to sample but it is feasible to obtain posterior samples from the full conditional distribution of each parameter. This algorithm generates a dependent sequence of parameter samples whose distribution converges to the joint posterior distribution. On the other hand, when both the joint posterior distribution and the full conditional distributions are not easy to sample, the Metropolis-Hastings (MH) algorithm allows us to generate a sequence of samples from the posterior distribution by generating a candidate sample from a proposal density given the previous sample, and accepting or rejecting a new candidate sample based on how probable the candidate sample is with respect to the previous sample.

However, in spatial applications, where the number of parameters is large, MCMC methods are computationally expensive. In consequence, other approaches have been proposed to overcome this problem. In particular, approximate Bayesian inference using integrated nested Laplace approximations (INLA) have shown better performance compared to MCMC methods (Rue et al., 2009) for latent Gaussian models.

2.5.1 INLA

The INLA method was proposed by Rue et al. (2009), it computes the marginal posterior density of the parameters through numerical approximations. For this to be achieved, it requires the model to belong to a particular type of latent Gaussian model (LGM).

Latent Gaussian Models

A latent Gaussian model is a Bayesian hierarchical model composed by three main levels. Let $\boldsymbol{Y}$ be a response variable and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ a random vector, then in general,

$$g(E(\boldsymbol{Y})) = \boldsymbol{Z}\boldsymbol{\beta} + f(.) + \epsilon,$$

where $g(.)$ is a link function, $\boldsymbol{Z}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of fixed effects, $f(.)$ are random effects, and ϵ is a random noise. Define $\boldsymbol{X} = \{\boldsymbol{\beta}, f(.), \epsilon\}$, and $\boldsymbol{\Psi}$ as the vector of hyperparameters. Then, the random vector $\boldsymbol{Y}$ lies in the first level and it is conditionally independent given $X = \boldsymbol{x}$ and $\boldsymbol{\Psi} = \boldsymbol{\psi}$. Therefore,

$$f_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\Psi}}(\boldsymbol{y}) = \prod_{i=1}^{n} f_{Y_i|\boldsymbol{X},\boldsymbol{\Psi}}(y_i).$$

The second level is composed of the latent vector $\boldsymbol{X}$ which, given $\boldsymbol{\Psi}$, has a Gaussian prior distribution $\boldsymbol{X}|\boldsymbol{\Psi} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}(\boldsymbol{\Psi}))$. This vector is called a latent Gaussian field and when its precision matrix $\boldsymbol{Q} = [\boldsymbol{\Sigma}(\boldsymbol{\Psi})]^{-1}$ is sparse, it is called a Gaussian Markov random field (GMRF). Finally, in the third level lies the vector of hyperparameters, $\boldsymbol{\Psi}$, which prior distribution does not depend on any other random variable. It is mainly composed of scale or dispersion parameters of $\boldsymbol{Y}$ and other parameters of the latent effects (precision parameters, generally).

Laplace approximation

In order to obtain the joint posterior distribution of $(\boldsymbol{X}, \boldsymbol{\Psi})$ given by

$$f(\boldsymbol{x}, \boldsymbol{\psi}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\psi}) f(\boldsymbol{x}|\boldsymbol{\psi}) f(\boldsymbol{\psi}).$$

INLA computes the marginal posterior density of the components of $\boldsymbol{X}$ and $\boldsymbol{\Psi}$ numerically integrating the following expressions:

$$f(x_j|\boldsymbol{y}) = \int f(x_j|\boldsymbol{\psi}, \boldsymbol{y}) f(\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi}, \quad j = 1, \ldots, J,$$

$$f(\psi_p|\boldsymbol{y}) = \int f(\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi_{-p}}, \quad p = 1, \ldots, P,$$

where x_j is the j-th component of vector $\boldsymbol{X}$, ψ_p is the p-th component of vector $\boldsymbol{\Psi}$, and $\boldsymbol{\psi_{-p}}$ is the vector $\boldsymbol{\Psi}$ without ψ_p. To solve these expressions, the first step is to find an approximation for

$$f(\boldsymbol{\psi}|\boldsymbol{y}) \propto \frac{f(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\psi})}{f(\boldsymbol{x}|\boldsymbol{\psi}, \boldsymbol{y})}.$$

Since $f(\boldsymbol{x}|\boldsymbol{\psi}, \boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\psi}) f(\boldsymbol{x}|\boldsymbol{\psi})$, then,

$$f(\boldsymbol{x}|\boldsymbol{\psi}, \boldsymbol{y}) \propto \exp(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \sum_{i=1}^{n} g_i(x_i)).$$

where $g_i(\boldsymbol{x}_i) = \log(f(y_i|\boldsymbol{x}_i, \boldsymbol{\psi}))$. Then, a Gaussian approximation for $f(\boldsymbol{x}|\boldsymbol{\psi}, \boldsymbol{y})$ of the form

$$\widetilde{f}_G(\boldsymbol{x}|\boldsymbol{\psi}, \boldsymbol{y}) \propto \exp(-\frac{1}{2}cx^2 + bx),$$

can be calculated using a quadratic Taylor expansion around $x^\star$ on $h(x) = -\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \sum_{i=1}^{n} g_i(\boldsymbol{x}_i)$, that is

$$h(x) \approx h(x^\star) + h'(x^\star)(x - x^\star) + \frac{1}{2}h''(x^\star)(x - x^\star)^2, \tag{2.11}$$

where $\boldsymbol{x}^\star$ is the mode of $f(\boldsymbol{x}|\boldsymbol{\psi}, \boldsymbol{y})$, which is found by means of the Newton-Raphson algorithm. From the equality of equation (2.11) and $a + bx - \frac{1}{2}cx^2$, it is obtained the Gaussian approximation $\widetilde{f}_G(\boldsymbol{x}|\boldsymbol{\psi}, \boldsymbol{y})$ which is the density function of $N(\boldsymbol{x}^\star, [\boldsymbol{Q} + diag(g_j''(x_j^\star))]^{-1})$. Finally, the Laplace approximation for $f(\boldsymbol{\psi}|\boldsymbol{y})$ is given by

$$\widetilde{f}(\boldsymbol{\psi}|\boldsymbol{y}) \propto \frac{f(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\psi})}{\widetilde{f}_G(\boldsymbol{x}|\boldsymbol{\psi}, \boldsymbol{y})}\Big|_{\boldsymbol{x}=\boldsymbol{x}^\star}.$$

The second step is to use $\widetilde{f}(\boldsymbol{\psi}|\boldsymbol{y})$ to calculate

$$f(\psi_p|\boldsymbol{y}) \approx \int \widetilde{f}(\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi_{-p}}; \quad p = 1, \ldots, P.$$

Then a set of suitable values of $\boldsymbol{\Psi}$ must be found for the numerical integration of $\widetilde{f}(\boldsymbol{\psi}|\boldsymbol{y})$. The exploration of $\widetilde{f}(\boldsymbol{\psi}|\boldsymbol{y})$ begins by localizing its mode, $\boldsymbol{\psi}^\star$. This is achieved through the optimization of $\log(\widetilde{f}(\boldsymbol{\psi}|\boldsymbol{y}))$. Then, a translation and rotation of coordinates from $\boldsymbol{\psi}$ to $\boldsymbol{z}$ is made, where the axes of $\boldsymbol{z}$ are the principal components of $\boldsymbol{\Sigma} = -\boldsymbol{H}^{-1}$ where $\boldsymbol{H}$ is the Hessian matrix at $\boldsymbol{\psi}^\star$. A grid of values of $\boldsymbol{\Psi}$ in $\boldsymbol{z}$ is chosen by localizing the points where the highest probability density is concentrated. This grid assigns more points to the highest

density regions. In order to return to the original coordinates, define $\boldsymbol{\psi}$ as a function of $\boldsymbol{z}$, as follows:

$$\boldsymbol{\psi}(\boldsymbol{z}) = \boldsymbol{\psi}^{\star} + \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{z},$$

where $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$ are matrices of the eigendecomposition $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\top}$ of $\boldsymbol{\Sigma}$. Finally, the marginal distribution of each ψ_p can be approximated using numerical integration, as follows:

$$\widetilde{f}(\psi_p|\boldsymbol{y}) = \sum_k \widetilde{f}(\psi^{(kp)}|\boldsymbol{y})w_{kp},$$

where k is the number of integration points, $\psi^{(kp)}$ is a point of $\boldsymbol{\Psi}$ where only the value of ψ_p remains constant for every k, and w_{kp} is its correspondent weight. The next step is to find an approximation for $f(x_j|\boldsymbol{\psi}, \boldsymbol{y})$. Since

$$f(x_j|\boldsymbol{\psi}, \boldsymbol{y}) \propto \frac{f(\boldsymbol{x}, \boldsymbol{\psi}|\boldsymbol{y})}{f(\boldsymbol{x}_{-j}|x_j, \boldsymbol{\psi}, \boldsymbol{y})}.$$

The Laplace approximation of $f(x_j|\boldsymbol{\psi}, \boldsymbol{y})$ is given by

$$\widetilde{f}(x_j|\boldsymbol{\psi}, \boldsymbol{y}) \approx \frac{f(\boldsymbol{x}, \boldsymbol{\psi}|\boldsymbol{y})}{\widetilde{f}_{GG}(\boldsymbol{x}_{-j}|x_j, \boldsymbol{\psi}, \boldsymbol{y})}\Big|_{\boldsymbol{x}_{-j}|\boldsymbol{x}^{\star}_{-j}(x_j,\boldsymbol{\psi})},$$

where $\widetilde{f}_{GG}$ is the Gaussian aproximation of $\widetilde{f}(\boldsymbol{x}_{-j}|x_j, \boldsymbol{\psi}, \boldsymbol{y})$. The final step is to use $\widetilde{f}(x_j|\boldsymbol{\psi}, \boldsymbol{y})$ and $\widetilde{f}(\boldsymbol{\psi}|\boldsymbol{y})$ to calculate

$$f(x_j|\boldsymbol{y}) \approx \int \widetilde{f}(x_j|\boldsymbol{\psi}, \boldsymbol{y})\widetilde{f}(\boldsymbol{\psi}|\boldsymbol{y})d\boldsymbol{\psi}; \quad j = 1, \ldots, J.$$

This approximation is done using the grid defined in the second step to numerically integrate

$$\widetilde{f}(x_j|\boldsymbol{y}) = \sum_k \widetilde{f}(x_j|\boldsymbol{\psi}^{(k)}, \boldsymbol{y})\widetilde{f}(\boldsymbol{\psi}^{(k)}|\boldsymbol{y})\Delta_k,$$

where $\boldsymbol{\psi}^{(\boldsymbol{k})} = (\psi_1^{(k)}, \ldots, \psi_P^{(k)})$ and Δ_k is its correspondent weight.

Chapter 3

Multivariate DAGAR model

Let assume that Y_{id} is a random variable representing the number of cases of response d in area i, that follows a Poisson distribution, such that:

$$Y_{id} \sim Poisson(\mu_{id} = E_{id}R_{id}); \quad i = 1, 2, \ldots, n; \quad d = 1, \ldots, K, \tag{3.1}$$

where μ_{id} represents the average number of cases of response d in the i-th area, E_{id} is called offset which represents the number of expected cases with response d in area i or possible cases with response d in area i, and R_{id} is the relative risk for each response d in area i. This relative risk is the modeled throughout:

$$\log(\mu_{id}) = \log(E_{id}) + \boldsymbol{z}_{id}^{\top}\boldsymbol{\beta_d} + \theta_{id},$$

where $\boldsymbol{z}_{id} = (1, z_{1id}, \ldots, z_{pid})$ is a vector of covariates, $\boldsymbol{\beta_d} = (\beta_{0d}, \ldots, \beta_{pd})$ is a vector of regression coefficients and θ_{id} represents the spatial multivariate random effect. The spatial random effects are represented by the Θ matrix with entries θ_{id} corresponding to the value of the response variable $d = 1, \ldots, K$ in the area $i = 1, \ldots, n$. Let $\Theta_{i\cdot}$ be the i-th row of the Θ matrix and $\Theta_{\cdot d}$ be the d-th column of the Θ matrix. The matrix $\boldsymbol{\Theta}$ is defined through the operator $vec(\cdot)$, such that:

$$vec(\Theta) = (\Theta_{\cdot 1}^{\top}, \ldots, \Theta_{\cdot K}^{\top})^{\top}.$$

Then the distribution of Θ is defined by:

$$vec(\Theta) \sim N(0, \mathbf{Q^{-1}} = \mathbf{\Gamma^{-1}} \otimes \mathbf{Q_S}^{-1}), \tag{3.2}$$

where $\mathbf{Q_S}$ is a spatial precision matrix of a DAGAR model and Γ controls the remaining variability between responses and $\mathbf{Q} = \mathbf{\Gamma} \otimes \mathbf{Q_S}$ is the precision matrix of $vec(\Theta)$.

Specifically, we assume that $\mathbf{\Gamma^{-1}}$ is defined as follows:

$$\mathbf{\Gamma^{-1}} = \begin{bmatrix} 1/\tau_1 & \rho_{12}/\sqrt{\tau_1\tau_2} & \cdots & \rho_{1K}/\sqrt{\tau_1\tau_K} \\ \rho_{12}/\sqrt{\tau_1\tau_2} & 1/\tau_2 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1K}/\sqrt{\tau_1\tau_K} & \cdots & \cdots & 1/\tau_K \end{bmatrix},$$

where ρ_{jl} is the correlation coefficient of j and l responses, and τ_d is the marginal precision of disease d. It is also assumed that $\Theta_{\cdot d} = (\theta_{1d}, \ldots, \theta_{nd}) \sim N(0, \mathbf{Q_S}^{-1})$ for any d,

$$\boldsymbol{Q}_S = \frac{1}{n!}\sum_{\pi} \boldsymbol{P}_\pi^T \boldsymbol{L}_\pi^T \boldsymbol{F}_\pi^{-1} \boldsymbol{L}_\pi \boldsymbol{P}_\pi, \tag{3.3}$$

where $\boldsymbol{L}_\pi = (\boldsymbol{I} - \boldsymbol{B}_\pi)$ and $\boldsymbol{P}_\pi$ is a permutation matrix corresponding to the ordering π. $\boldsymbol{Q}$ is a sparse matrix since $\boldsymbol{Q}_S$ is sparse. The components of $\boldsymbol{B}_\pi$ and $\boldsymbol{F}_\pi$ are defined in equation 2.10, where ρ from now on will be referred to as α, so that:

$$b_{ij} = \frac{\alpha}{1 + (n_{\pi_{(i)}} - 1)\alpha^2}; \quad \tau_i = (\sigma_i^2)^{-1} = \frac{1 + (n_{\pi(i)} - 1)\alpha^2}{1 - \alpha^2},$$

where $n_{\pi(i)}$ denotes the cardinality of $N_\pi(i)$, that is, the number of neighbors of an area i.

3.1 Bayesian inference for multivariate DAGAR model

In order to estimate the proposed model using INLA, we redefine the model as a latent Gaussian model. First, we assume that $Y'_{id}s$ are conditional independent given the latent Gaussian field $\boldsymbol{X}$, and the vector of hyperparameters $\mathbf{\Psi}$, that is,

$$Y_{id}|\boldsymbol{X}, \mathbf{\Psi} \overset{ind}{\sim} Poisson(\mu_{id} = E_{id}R_{id}).$$

Then,

$$f(\boldsymbol{Y}|\boldsymbol{X}, \mathbf{\Psi}) = \prod_{d=1}^{K}\prod_{i=1}^{n} f_{Y_{id}|X,\Psi}(y_{id}). \tag{3.4}$$

Second, let define the vector $\boldsymbol{X} = (\boldsymbol{\beta}, vec(\mathbf{\Theta}))$, where $\boldsymbol{\beta} = (\boldsymbol{\beta_1^\top}, \ldots, \boldsymbol{\beta_K^\top})$. Let also assume that the components of $\boldsymbol{\beta_d}$ for $d = 1, \ldots, K$, that is β_{ld} for $l = 0, \ldots, p$, are normally distributed and independent, that is, $\beta_{ld} \overset{ind}{\sim} N(0, \sigma_\beta^2)$, consequently:

$$f(\boldsymbol{\beta}) = \prod_{d=1}^{K}\prod_{l=0}^{p} f_{\beta_{ld}}(\beta_{ld}) \tag{3.5}$$

Then, $\boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_\beta)$, where $\Sigma_\beta = diag(\sigma_\beta^2)$ is square matrix of dimension $(p+1)d$. Also, assume that $vec(\boldsymbol{\Theta})$ is normally distributed, as stated in equation (3.2), where $\boldsymbol{Q}$ is a function of $\boldsymbol{\Psi}$ and that $\boldsymbol{\beta}$ and $vec(\boldsymbol{\Theta})$ are independent. Therefore,

$$\boldsymbol{X}|\boldsymbol{\Psi} = (\boldsymbol{\beta}, vec(\boldsymbol{\Theta}))|\boldsymbol{\Psi} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_X = \begin{bmatrix} \boldsymbol{\Sigma}_\beta & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}^{-1}(\boldsymbol{\Psi}) \end{bmatrix}\right). \tag{3.6}$$

Because $\boldsymbol{Q}(\boldsymbol{\Psi})$ is sparse ($\boldsymbol{Q}_S$ is sparse) and $\boldsymbol{\Sigma}_\beta^{-1}$ is a diagonal matrix, then

$$\boldsymbol{Q}_X(\boldsymbol{\Psi}) = \boldsymbol{\Sigma}_X^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_\beta^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}(\boldsymbol{\Psi}). \end{bmatrix}$$

Note that $\boldsymbol{Q}_X(\boldsymbol{\Psi})$ is also a sparse matrix, which makes $\boldsymbol{X}$ a GMRF. Finally, the vector of hyperparameters $\boldsymbol{\Psi}$ is composed of $\boldsymbol{\Gamma}$ and the parameter involved in $\boldsymbol{Q}_S$, that is $\boldsymbol{\Psi} = (\boldsymbol{\Gamma}, \alpha)$, with prior distributions:

$$\boldsymbol{\Gamma} \sim Wishart(r, \boldsymbol{R}^{-1}),$$
$$\alpha \sim U(0,1),$$

where r represents the degrees of freedom, $\boldsymbol{R}^{-1}$ is a scale matrix. Since $\boldsymbol{\Gamma}$ and α are independent, then $f(\boldsymbol{\Psi}) = f(\boldsymbol{\Gamma})f(\alpha)$.

Now that the model meets the requirement to be estimated by INLA, the estimation of the marginal distributions of the components of $\boldsymbol{X}$ and $\boldsymbol{\Psi}$ will be obtained by numerically integrating

$$f(x_j|\boldsymbol{y}) = \int f(x_j|\boldsymbol{\psi}, \boldsymbol{y}) f(\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi}, \quad j = 1, \ldots, J,$$

$$f(\psi_p|\boldsymbol{y}) = \int f(\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi_{-p}}, \quad p = 1, \ldots, P.$$

This is achieved by finding the Laplace approximation of $f(\boldsymbol{\psi}|\boldsymbol{y})$ and $f(x_j|\boldsymbol{\psi}, \boldsymbol{y})$, given by

$$\widetilde{f}(\boldsymbol{\psi}|\boldsymbol{y}) \propto \frac{f(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\psi})}{\widetilde{f}_G(\boldsymbol{x}|\boldsymbol{\psi}, \boldsymbol{y})}\Big|_{\boldsymbol{x}=\boldsymbol{x}^\star},$$

$$\widetilde{f}(x_j|\boldsymbol{\psi}, \boldsymbol{y}) \approx \frac{f(\boldsymbol{x}, \boldsymbol{\psi}|\boldsymbol{y})}{\widetilde{f}_{GG}(\boldsymbol{x}_{-j}|x_j, \boldsymbol{\psi}, \boldsymbol{y})}\Big|_{\boldsymbol{x}_{-j}=\boldsymbol{x}^\star_{-j}(x_j,\boldsymbol{\psi})},$$

where $\boldsymbol{x}^\star$ and $\boldsymbol{x}^\star_{-j}(x_j, \boldsymbol{\psi})$ are the modes of the distributions of $\boldsymbol{x}|\boldsymbol{\psi}, \boldsymbol{y})$ and $\boldsymbol{x}_{-j}|x_j, \boldsymbol{\psi}, \boldsymbol{y}$, respectively. Then the marginal distributions are computed using these approximations:

$$f(\psi_p|\boldsymbol{y}) \approx \int \widetilde{f}(\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi_{-p}}, \quad p = 1, \ldots, P,$$

$$f(x_j|\boldsymbol{y}) \approx \int \widetilde{f}(x_j|\boldsymbol{\psi}, \boldsymbol{y})\widetilde{f}(\boldsymbol{\psi}|\boldsymbol{y})d\boldsymbol{\psi}, \quad j = 1, \ldots, J,$$

through numerical integration.

Chapter 4

Simulation study

In order to study the performance of the multivariate DAGAR model, simulations studies under multiple scenarios are presented in this chapter.

The simulations were performed for $K = 3$ three response variables using the map of the provinces of Peru (196 provinces). First, it is simulated the spatial multivariate random effect θ_{id} of a response (disease) d in the i-th province of Peru, for $d = 1, 2, 3$ and $i = 1, \ldots, 196$. In particular, $\boldsymbol{\theta_d} = (\theta_{1d}, \ldots, \theta_{196d})$, is a random vector that contains the spatial multivariate random effects of each province for disease d. Then the vector of spatial multivariate random effects $vec(\Theta) = (\boldsymbol{\theta_1}^\top, \boldsymbol{\theta_2}^\top, \boldsymbol{\theta_3}^\top)$ were simulated from a normal distribution according to:

$$vec(\Theta) \sim N(0, \mathbf{Q^{-1}} = \mathbf{\Gamma^{-1}} \otimes \mathbf{Q_S}^{-1}),$$

where $\mathbf{\Gamma^{-1}}$ is a precision matrix that takes into account the autocorrelation between the response variables and $\mathbf{Q_S}$ is a precision matrix to model the spatial autocorrelation between provinces. Specifically, $\mathbf{\Gamma^{-1}}$ is defined as follows:

$$\mathbf{\Gamma^{-1}} = \begin{bmatrix} 1/\tau_1 & \rho_{12}/\sqrt{\tau_1\tau_2} & \rho_{13}/\sqrt{\tau_1\tau_3} \\ \rho_{12}/\sqrt{\tau_1\tau_2} & 1/\tau_2 & \rho_{23}/\sqrt{\tau_2\tau_3} \\ \rho_{13}/\sqrt{\tau_1\tau_3} & \rho_{23}/\sqrt{\tau_2\tau_3} & 1/\tau_3 \end{bmatrix},$$

where ρ_{12} is the correlation coefficient of 1 and 2 responses, ρ_{13} is the correlation coefficient of 1 and 3 responses, and ρ_{23} is the correlation coefficient of 2 and 3 responses, and τ_d is the marginal precision of response d. On the other hand, the form of $\mathbf{Q_S}$ depends on the type of spatial multivariate random effect being simulated, we consider three types:

i) For MCAR spatial multivariate random effects, $\boldsymbol{Q_S} = \frac{1}{\sigma^2}(\boldsymbol{D_w} - \alpha \boldsymbol{W})$, as defined in Section 2.4.1.

ii) For MDAGAR spatial multivariate random effects, $\boldsymbol{Q}_S = \frac{1}{n!}\sum_\pi \boldsymbol{P}_\pi^T \boldsymbol{L}_\pi^T \boldsymbol{F}_\pi^{-1} \boldsymbol{L}_\pi \boldsymbol{P}_\pi$, as defined in equation (3.3), where $\boldsymbol{P}_\pi$ is a permutation matrix corresponding to the ordering π, $\boldsymbol{L}_\pi = (\boldsymbol{I} - \boldsymbol{B}_\pi)$, and the components of $\boldsymbol{B}_\pi$ and $\boldsymbol{F}_\pi$ are:

$$b_{ij} = \frac{\alpha}{1 + (n_{\pi_{(i)}} - 1)\alpha^2}; \quad \tau_i = (\sigma_i^2)^{-1} = \frac{1 + (n_{\pi(i)} - 1)\alpha^2}{1 - \alpha^2},$$

where $n_{\pi(i)}$ denotes the cardinality of $N_\pi(i)$, that is, the number of neighbors of an area i.

iii) For exponential Gaussian process spatial multivariate random effects, $\boldsymbol{Q_S = G^{-1}}$, for $G(d(i,j)) = \exp(-\phi d(i,j))$, where $\phi = -\log(\alpha)$ and $d(i,j)$ is the distance between the i-th and j-th province after embedding the graph of the centroids of the provinces of Peru in a 196×196 grid in the Euclidean plane and scaling the resulting distance matrix so that the mean distance between neighbors is one.

Then the average number of cases of response d in the i-th province is computed from $\mu_{id} = \exp(\beta_{0d} + \beta_{1d} x_i + \theta_{id})$, where β_{0d} and β_{1d} are the regression coefficients, $x_i \sim N(0,1)$ is the value of the covariate x in the i-th province. Finally, the response variables are simulated from $Y_{id} \overset{ind}{\sim} Poisson(\mu_{id})$.

For all scenarios presented, regression coefficients $\beta_{01} = 2$, $\beta_{11} = 0.6$, $\beta_{02} = 2.5$, $\beta_{12} =$ -0.1, $\beta_{03} = 3$, and $\beta_{13} = 0.2$ were considered. Furthermore, the hyperparameters values were set as follows, $\rho_{12} = 0.8$, $\rho_{13} = 0.6$, and $\rho_{23} = 0.4$, $\tau_1 = 2$, $\tau_2 = 1$, and $\tau_3 = 1.5$. We also consider different scenarios for the value of spatial autocorrelation α.

First, we study the right implementation of the MCAR and DAGAR models, simulating from each model and fitting the corresponding model. Therefore, in the first simulation, we simulate data from MCAR model and fit the MCAR model. In the second simulation, we simulate data from MDAGAR model and fit the MDAGAR model. In the third and fourth simulation, we compare the performance of the MCAR and MDAGAR models through replicated simulations.

4.1 Simulation 1: MCAR models

Areal data were simulated using MCAR model for different scenarios with $\alpha = 0.4$, 0.6, and 0.9. Then, a MCAR model was fitted. The simulated data and the results from each model are presented below.

Figure 4.1 shows in the first row histograms of the data simulated from a Poisson distribu-

tion with the aforementioned characteristics and $\alpha = 0.4$, the second row shows the simulated values of the response variables and the third row shows the spatial random effects in the region under study, respectively. Similar plots are presented in Figure 4.2 and Figure 4.3, for $\alpha = 0.6$ and $\alpha = 0.9$, respectively.

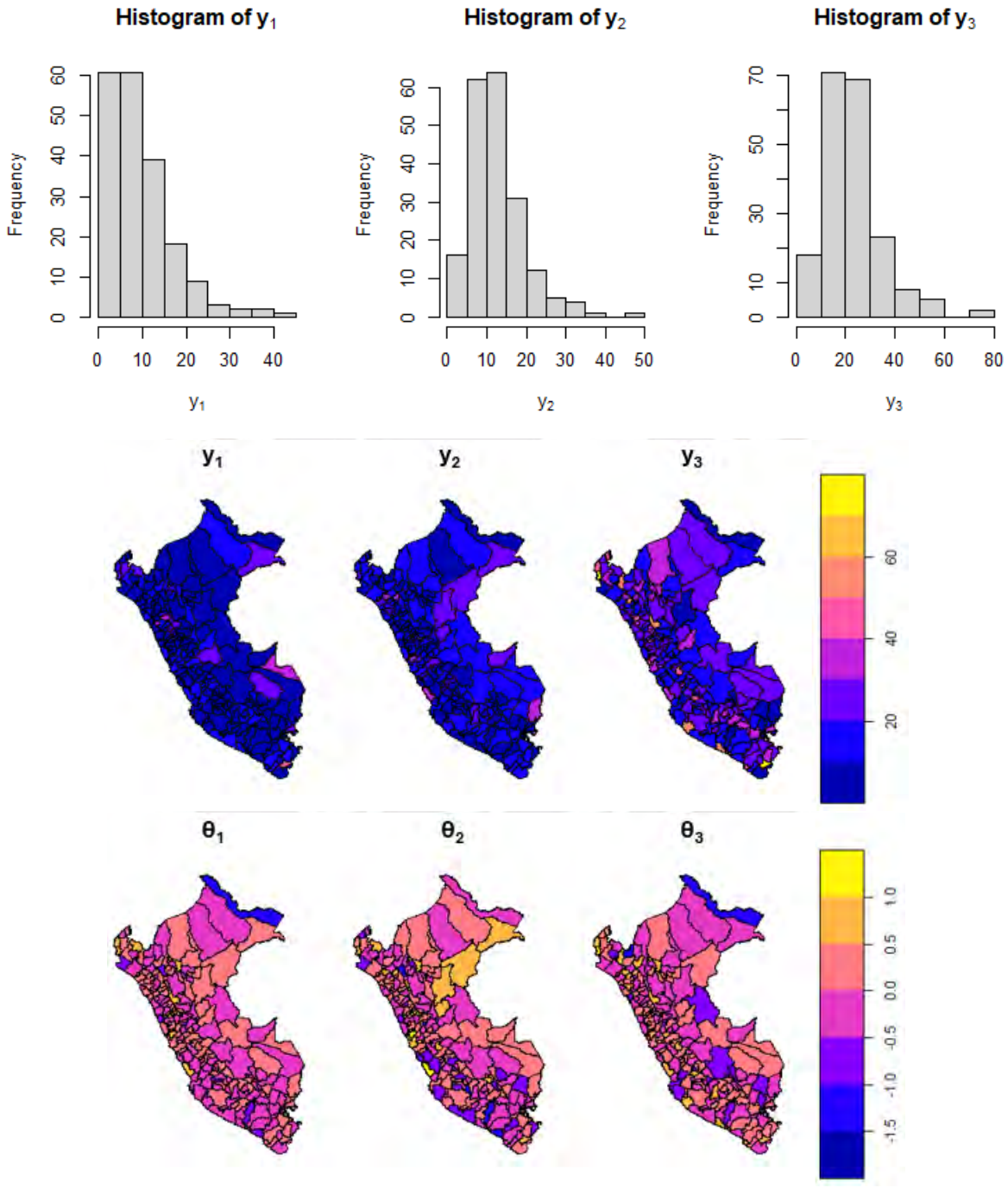


Figure 4.1: First row: Histograms of the three response variables obtained by simulation with MCAR random effects and $\alpha = 0.4$. Second row: Maps of the three response variables obtained by simulation with MCAR random effects. Third row: Maps of the simulated MCAR random effects corresponding to the three response variables.

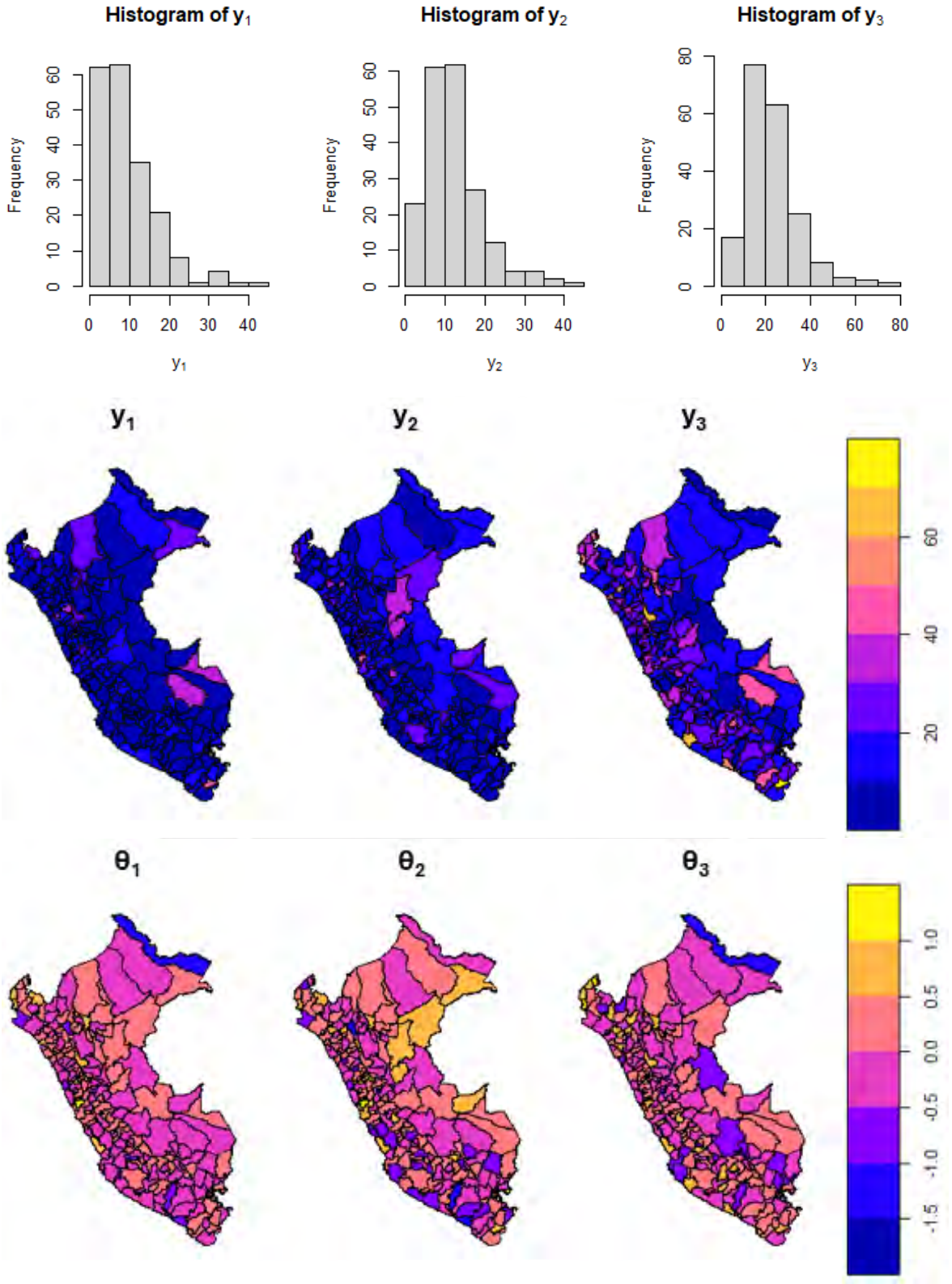


Figure 4.2: First row: Histograms of the three response variables obtained by simulation with MCAR random effects and $\alpha = 0.6$. Second row: Maps of the three response variables obtained by simulation with MCAR random effects. Third row: Maps of the simulated MCAR random effects corresponding to the three response variables.

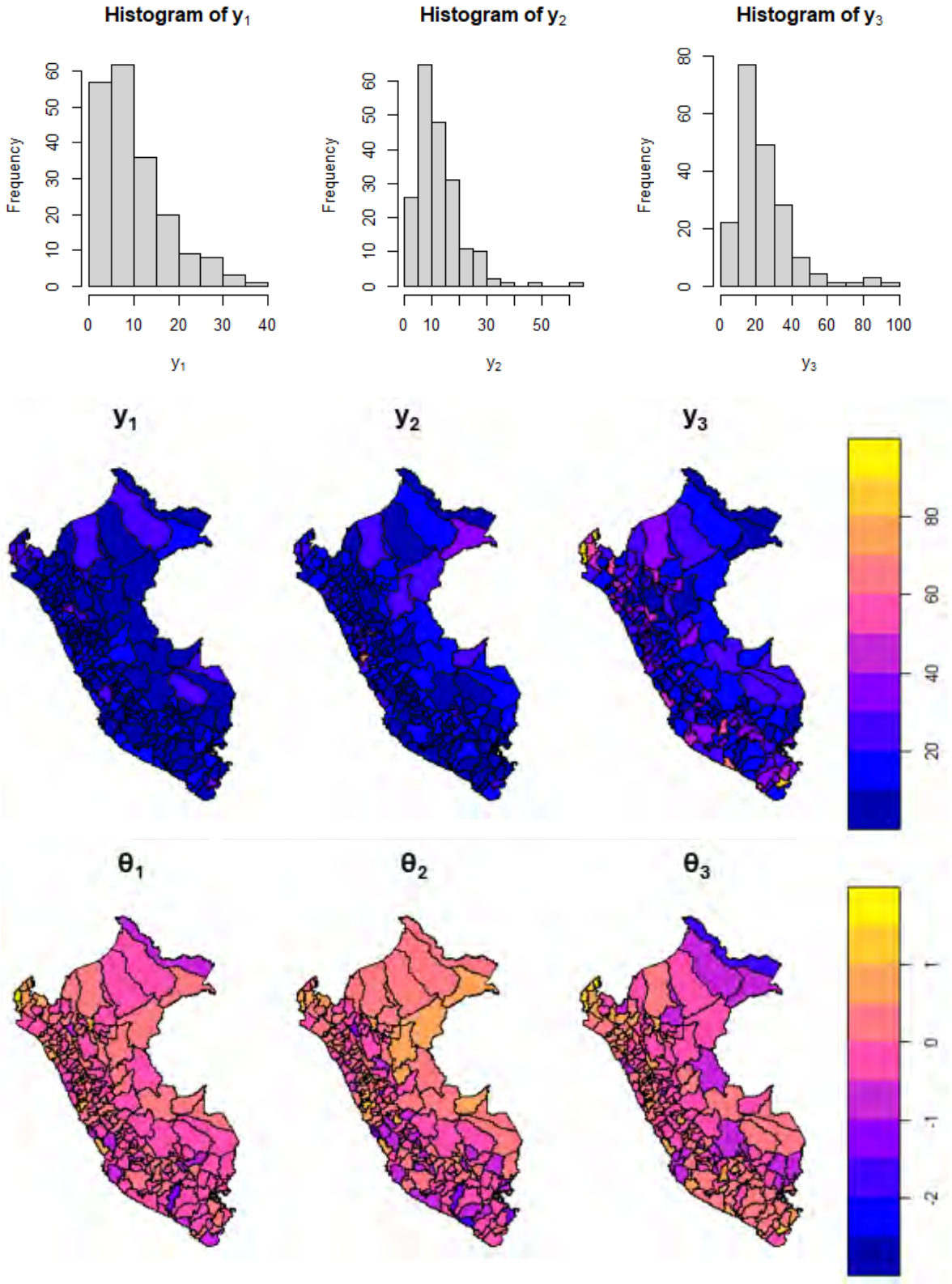


Figure 4.3: First row: Histograms of the three response variables obtained by simulation with MCAR random effects and $\alpha = 0.9$. Second row: Maps of the three response variables obtained by simulation with MCAR random effects. Third row: Maps of the simulated MCAR random effects corresponding to the three response variables.

Table 4.1 shows the mean, standard deviation, and credible intervals (95%) of the fixed effects and hyperparameters estimated by the MCAR models.

Table 4.1: Mean, standard deviation (SD), Lower Limit (LL) and Upper Limit (UL) of 95% credible intervals for the MCAR models with $\alpha = 0.4$, 0.6 and 0.9.

Response	Parameter	True value	Mean	SD	LL (95%)	UL (95%)
$\boldsymbol{\alpha = 0.4}$						
y_1	β_{01}	2	1.995	0.04	1.916	2.072
	β_{11}	0.6	0.622	0.037	0.55	0.694
	τ_1	2	2.226	0.437	1.492	3.204
y_2	β_{02}	2.5	2.505	0.039	2.427	2.581
	β_{12}	-0.1	-0.057	0.036	-0.128	0.013
	τ_2	1	1.361	0.211	0.993	1.819
y_3	β_{03}	3	3.019	0.036	2.948	3.088
	β_{13}	0.2	0.222	0.032	0.159	0.286
	τ_3	1.5	1.483	0.201	1.128	1.915
	ρ_{12}	0.8	0.647	0.087	0.455	0.794
	ρ_{13}	0.6	0.449	0.097	0.244	0.624
	ρ_{23}	0.4	0.204	0.095	0.014	0.387
	α	0.4	0.257	0.13	0.059	0.545
$\alpha = \mathbf{0.6}$						
y_1	β_{01}	2	1.98	0.044	1.893	2.065
	β_{11}	0.6	0.627	0.036	0.556	0.698
	τ_1	2	2.491	0.527	1.617	3.681
y_2	β_{02}	2.5	2.452	0.052	2.35	2.553
	β_{12}	-0.1	-0.09	0.039	-0.166	-0.014
	τ_2	1	1.117	0.17	0.818	1.486
y_3	β_{03}	3	3.008	0.04	2.928	3.087
	β_{13}	0.2	0.238	0.03	0.179	0.297
	τ_3	1.5	1.833	0.265	1.368	2.405
	ρ_{12}	0.8	0.613	0.096	0.403	0.776
	ρ_{13}	0.6	0.479	0.1	0.266	0.656
	ρ_{23}	0.4	0.39	0.089	0.206	0.556
	α	0.6	0.544	0.127	0.288	0.774
$\alpha = \mathbf{0.9}$						
y_1	β_{01}	2	2.014	0.072	1.87	2.155
	β_{11}	0.6	0.63	0.037	0.559	0.703
	τ_1	2	2.351	0.437	1.607	3.318
y_2	β_{02}	2.5	2.437	0.093	2.251	2.621
	β_{12}	-0.1	-0.036	0.038	-0.112	0.039
	τ_2	1	1.219	0.185	0.897	1.622
y_3	β_{03}	3	2.994	0.086	2.821	3.164
	β_{13}	0.2	0.218	0.034	0.152	0.283
	τ_3	1.5	1.401	0.195	1.059	1.823
	ρ_{12}	0.8	0.681	0.075	0.516	0.807
	ρ_{13}	0.6	0.554	0.085	0.372	0.703
	ρ_{23}	0.4	0.226	0.092	0.043	0.401
	α	0.9	0.882	0.047	0.773	0.954

From Table 4.1, when α =0.4, the model does a good job estimating the parameters, but it is more difficult to estimate ρ_{12} and ρ_{23}, which credible intervals do not include the true values. This can also be observed in the marginal posterior distributions showed in Figure 4.4. From Table 4.1, when α =0.6, the model does a good job estimating the parameters, but it is more difficult to estimate ρ_{12} which credible interval does not include the true value. This can also be observed in in the marginal posterior distributions showed in Figure 4.5.From Table 4.1, when α =0.9, the model does a good job estimating all the parameters. This can also be observed in the marginal posterior distributions showed in the third and forth rows of Figure 4.5. This indicates an improve of performance of the MCAR model estimating the hyperparameters when fitting multivariate areal data with high values of spatial autocorrelation α.

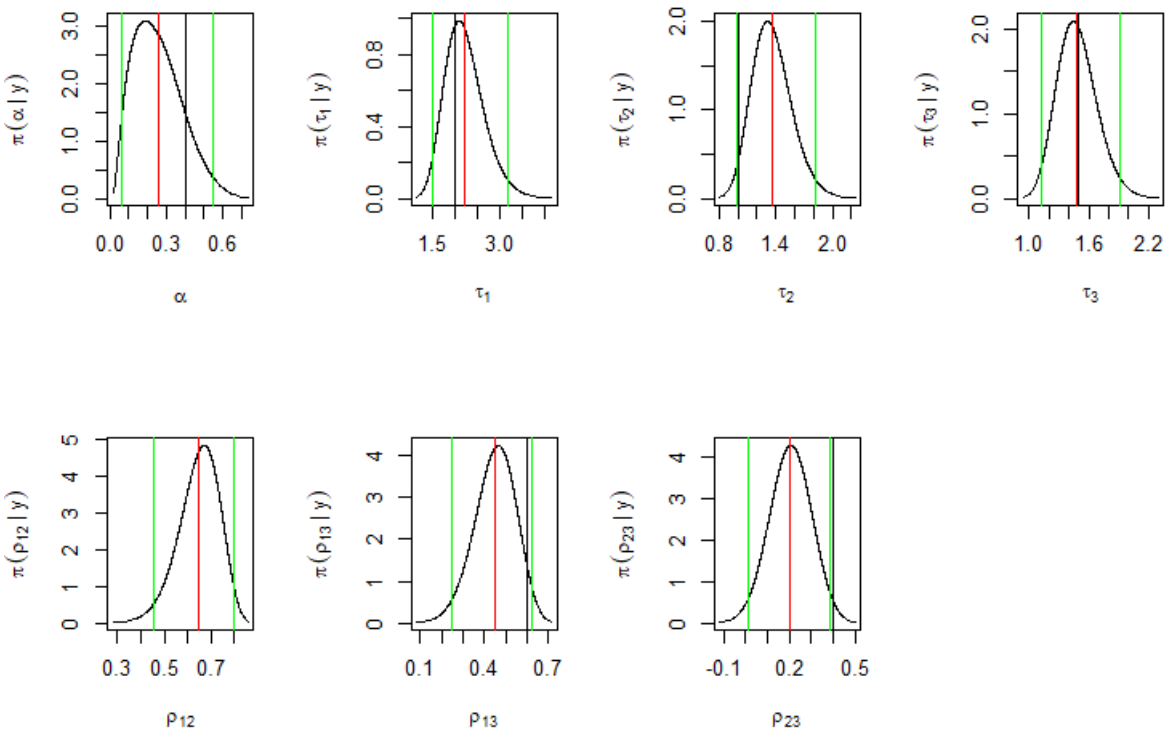


Figure 4.4: Marginal posterior density function of the hyperparameters estimated using the MCAR model for α = 0.1. The black vertical line indicates the true value of the hyperparameter. The red line indicates the mean of the estimated distribution. The green lines indicates the lower and upper limits of the 95% credible interval.

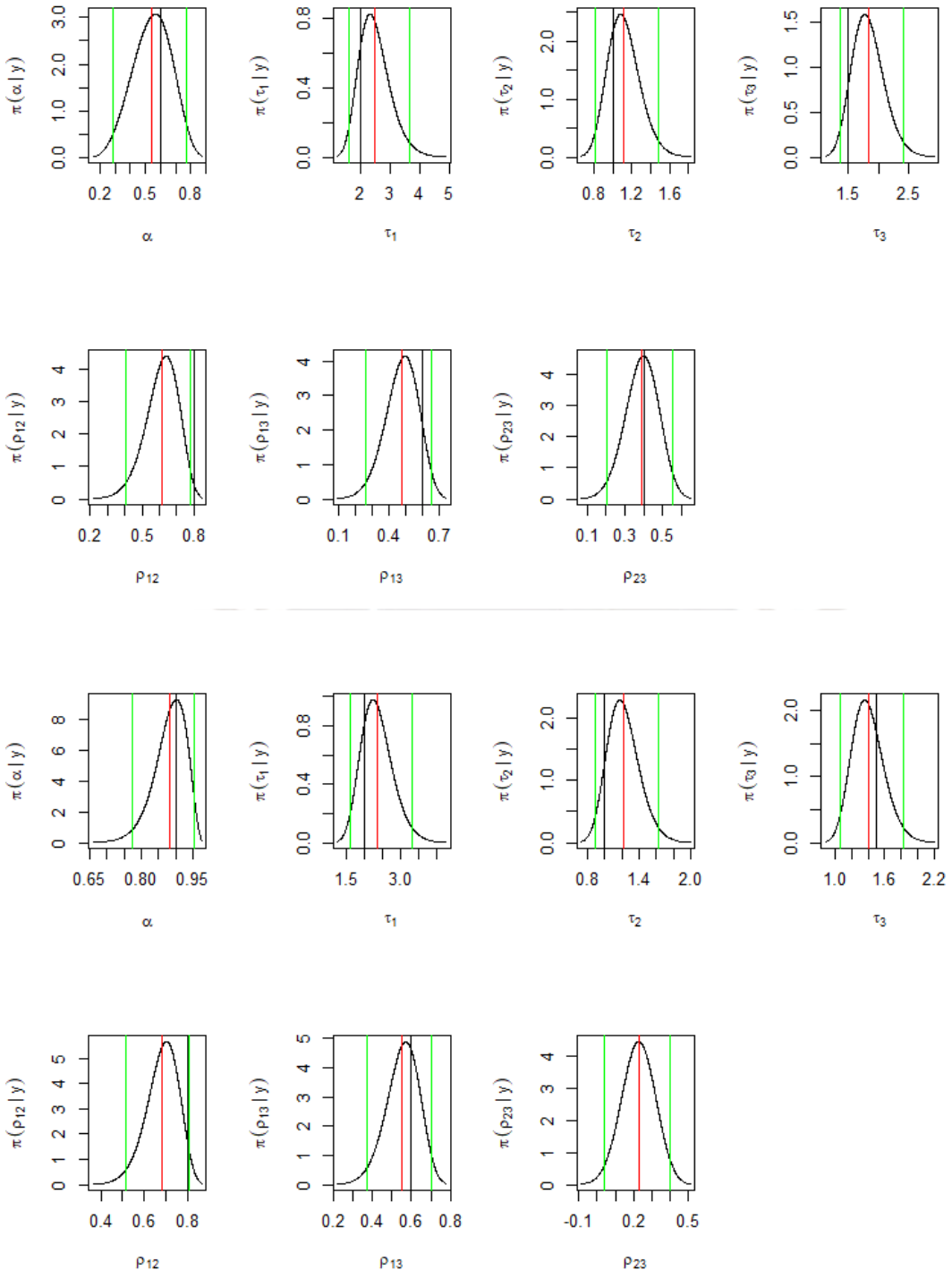


Figure 4.5: Marginal posterior density function of the hyperparameters estimated using the MCAR model. The black vertical line indicates the true value of the hyperparameter. The red line indicates the mean of the estimated distribution. The green lines indicates the lower and upper limits of the 95% credible interval. For α =0.6 (first and second row) and α =0.9 (third and fourth row).

The spatial random effects and the response variables are also well estimated as can be seen in Figures 4.6 and 4.7, respectively. In Figure 4.6 we can see that the estimates of the spatial random effects θ_{id} tend to be closer to the true value as α increases. In Figure 4.7 it can be observed that the estimations of the third response variable Y_3 are better than those of the other two response variables Y_1 and Y_2, for any value of α. This means that the estimation of the response variable is also better for higher values of the spatial autocorrelation α.

In summary, the MCAR model does a good job estimating the parameters of the simulated model, performing better for higher values of α. This results are only for one simulation. For a more accurate assessment of the model estimating capacity, more simulations are needed. These results are shown in Section 4.3.

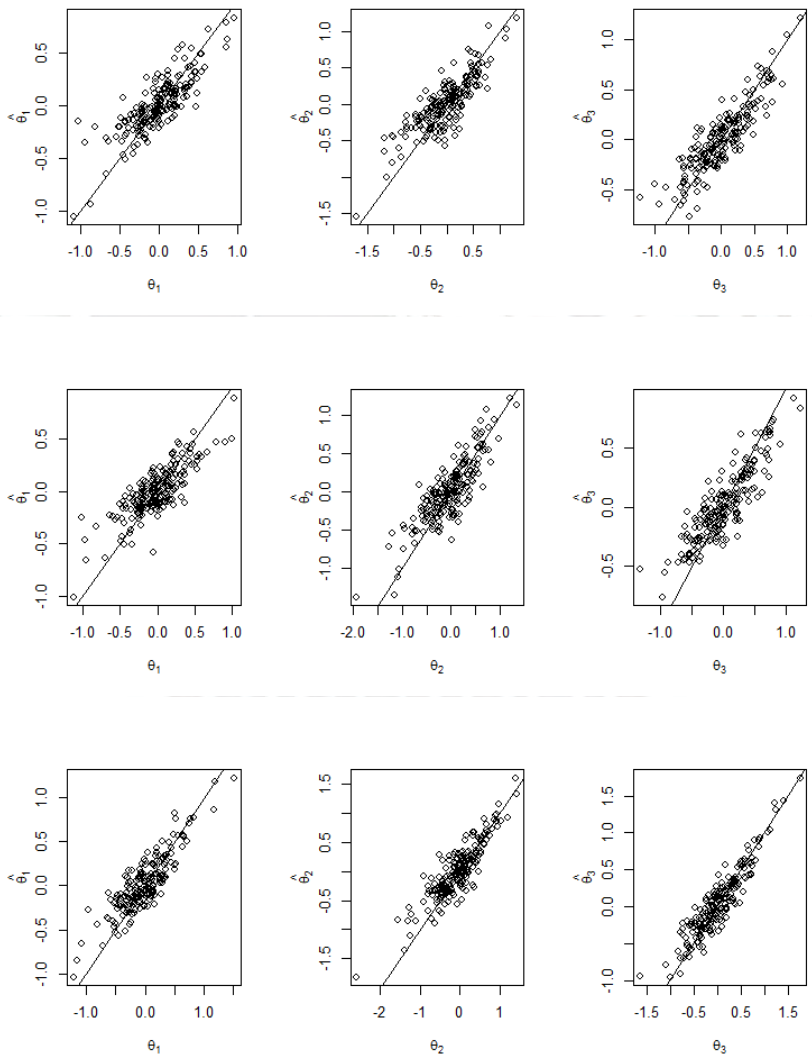


Figure 4.6: Plots of the spatial random effects corresponding to the three response variables against their estimations from the MCAR model. A line of slope equal to 1 is plotted for reference. For α =0.4 (first row), α =0.6 (second row) and α =0.9 (third row).

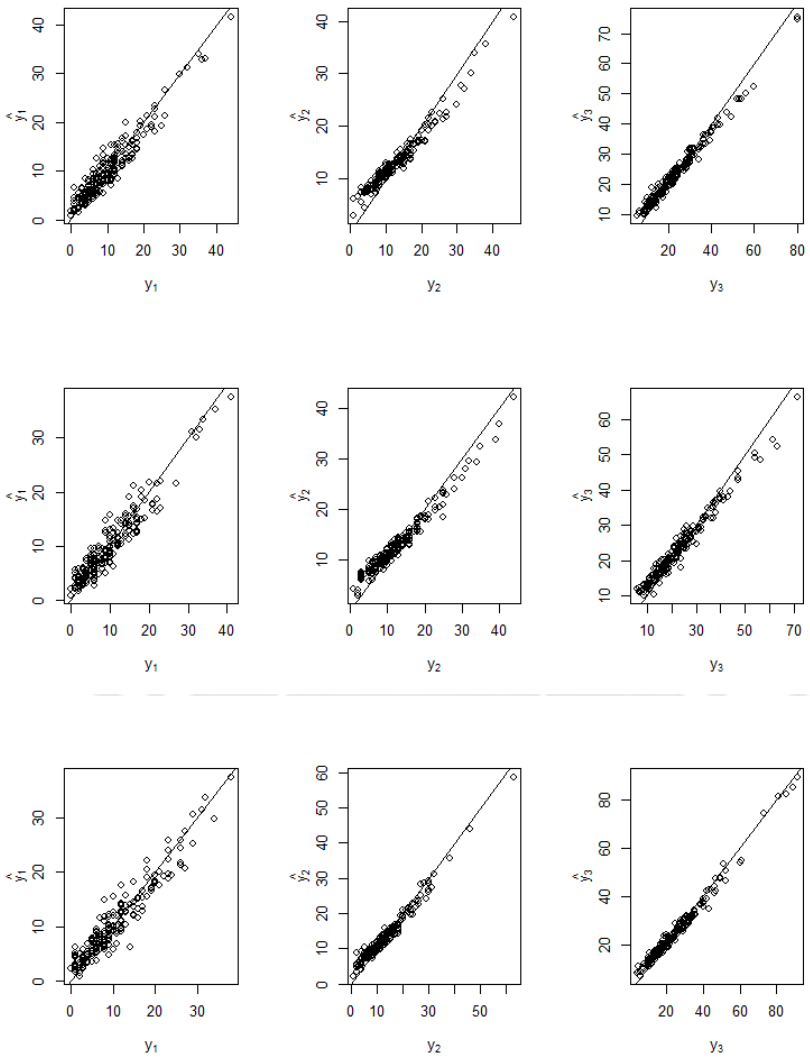


Figure 4.7: Plots of the three response variables simulated against their estimations from the MCAR model. A line of slope equal to 1 is plotted for reference. For α =0.4 (first row), α =0.6 (second row) and α =0.9 (third row).

4.2 Simulation 2: MDAGAR models

In this simulation, areal data were simulated using MDAGAR random effects for $\alpha = 0.4$, 0.6, and 0.9. Then, a MDAGAR model was fitted.

Figure 4.8 shows in the first row histograms of the data simulated from a Poisson distribution with the aforementioned characteristics and $\alpha = 0.4$, the second row shows the simulated values of the response variables and the third row shows the spatial random effects in the region under study, respectively. Similar plots are presented in Figure 4.9 and Figure 4.10, for $\alpha = 0.6$ and $\alpha = 0.9$, respectively.

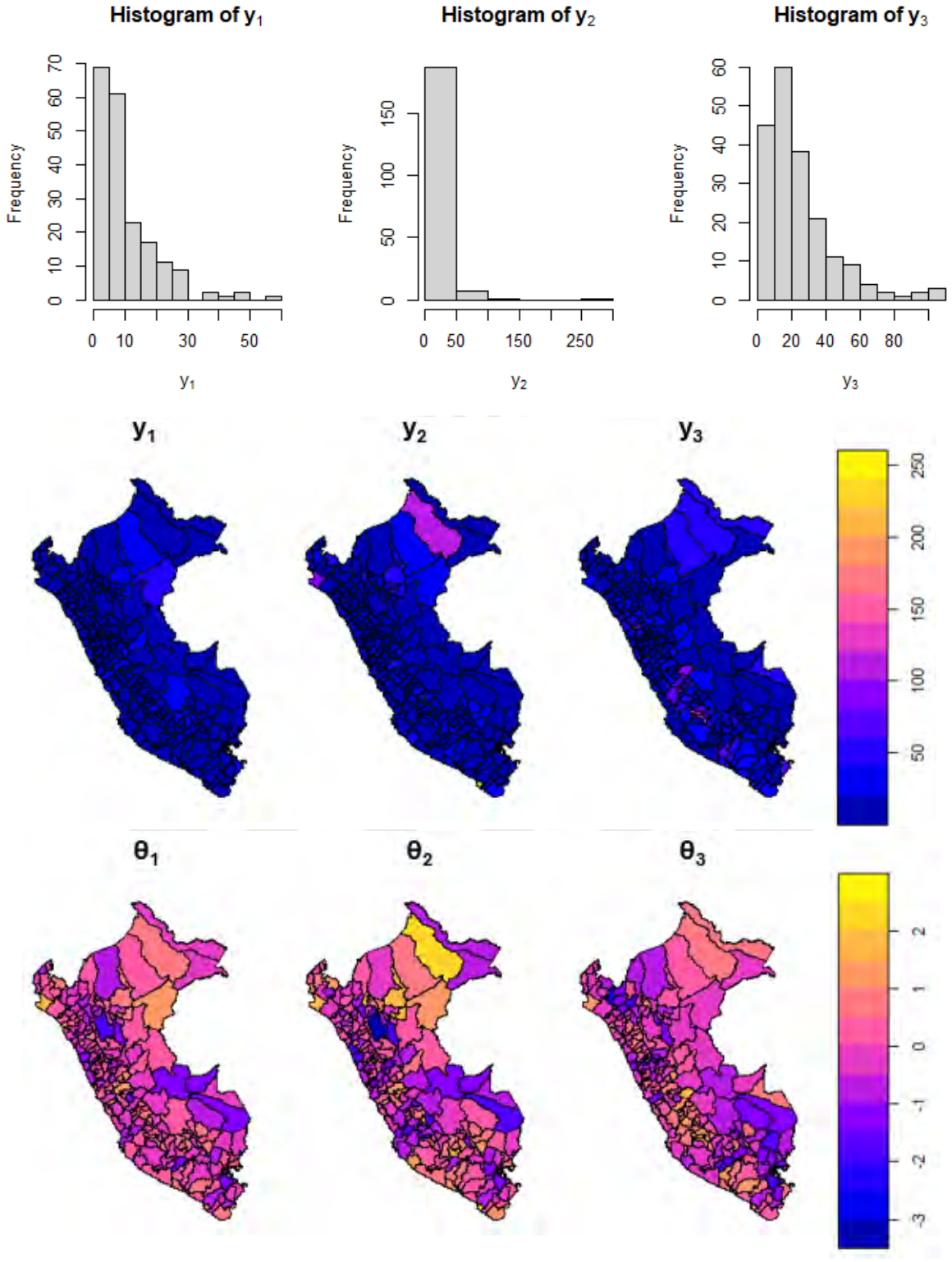


Figure 4.8: First row: Histograms of the three response variables obtained by simulation with MDAGAR random effects and $\alpha = 0.4$. Second row: Maps of the three response variables obtained by simulation with MDAGAR random effects. Third row: Maps of the simulated MDAGAR random effects corresponding to the three response variables.

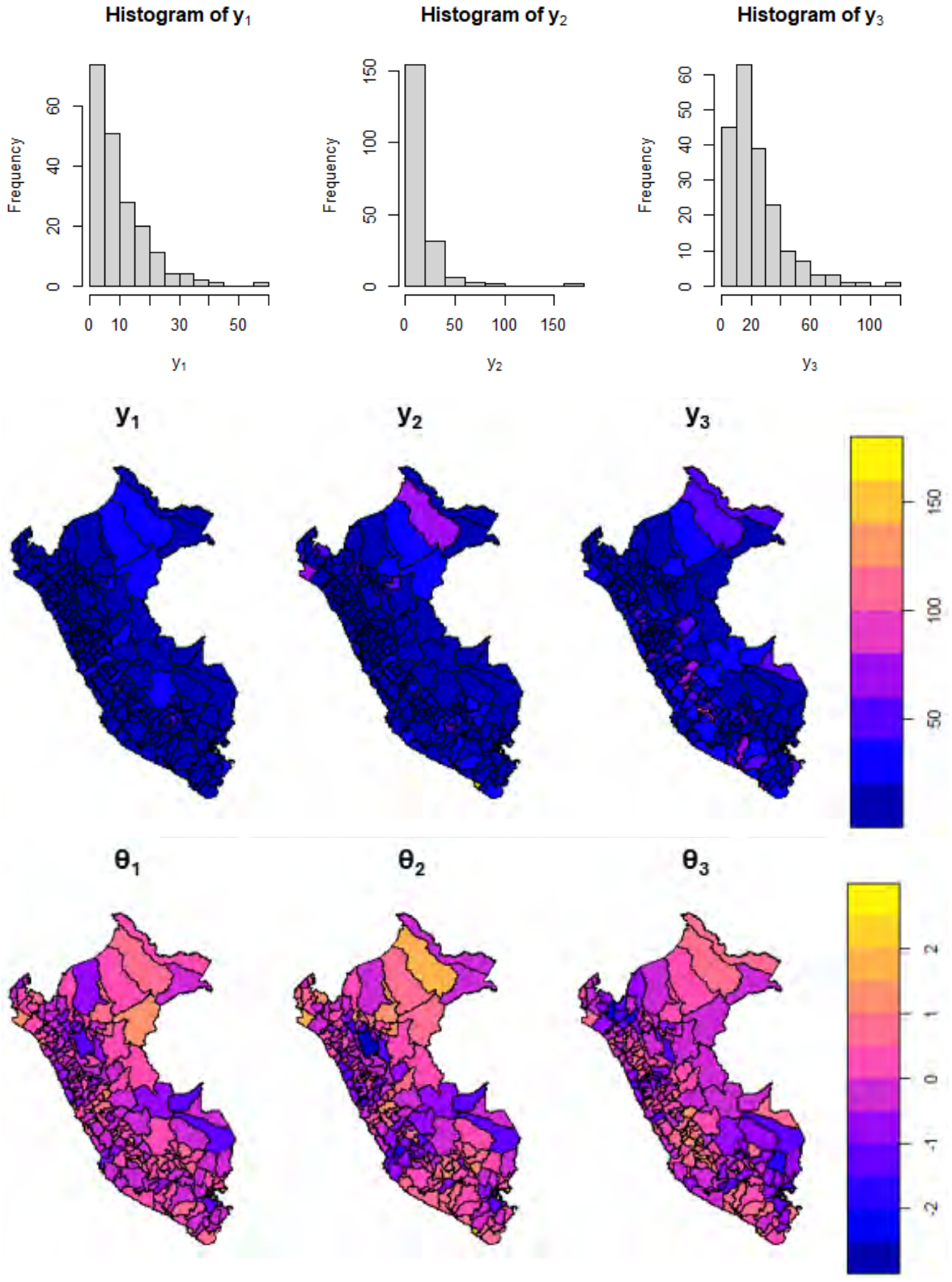


Figure 4.9: First row: Histograms of the three response variables obtained by simulation with MDAGAR random effects and $\alpha = 0.6$. Second row: Maps of the three response variables obtained by simulation with MDAGAR random effects. Third row: Maps of the simulated MDAGAR random effects corresponding to the three response variables.

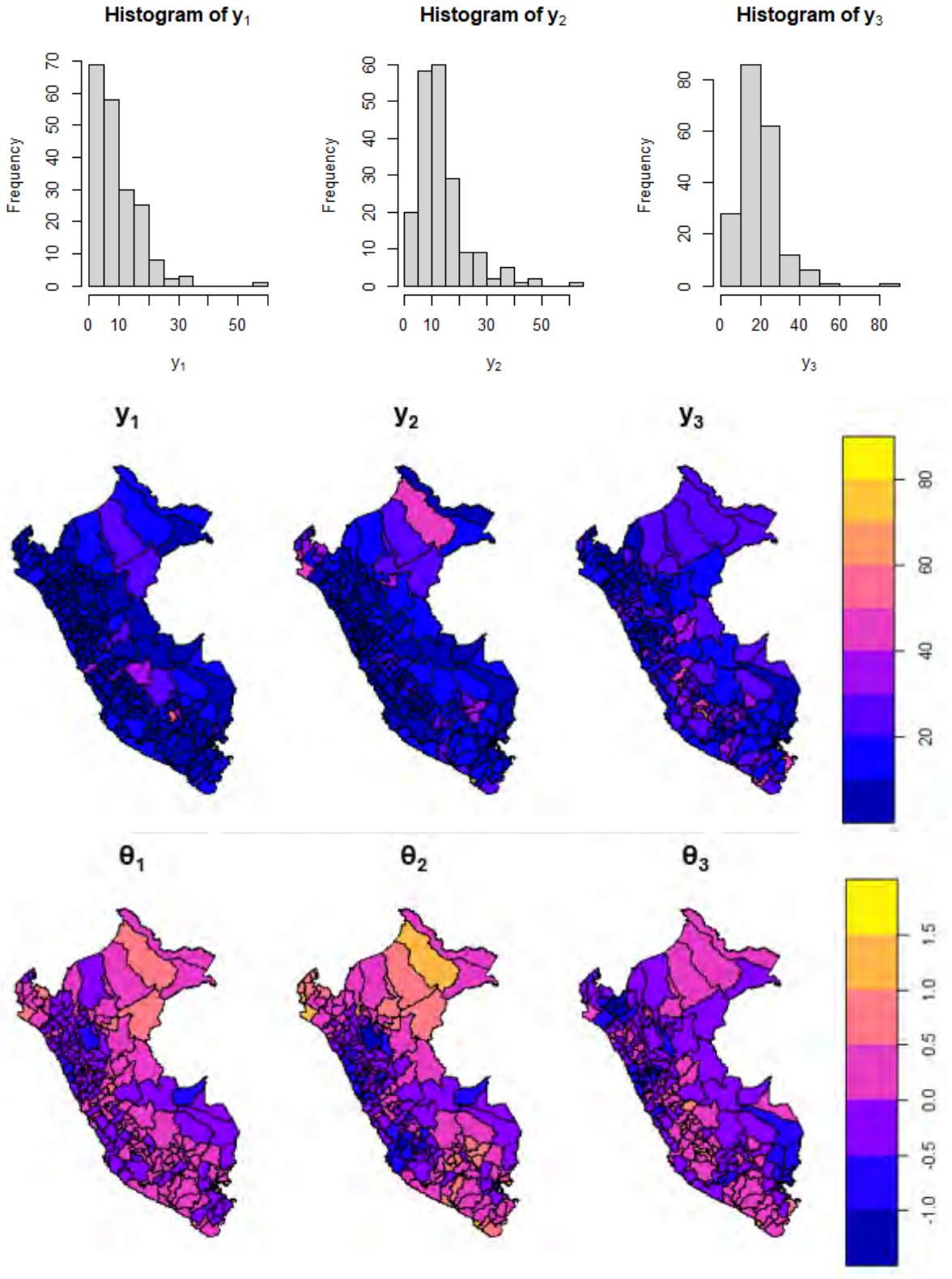


Figure 4.10: First row: Histograms of the three response variables obtained by simulation with MDAGAR random effects and $\alpha = 0.9$. Second row: Maps of the three response variables obtained by simulation with MDAGAR random effects. Third row: Maps of the simulated MDAGAR random effects corresponding to the three response variables.

Table 4.2 shows the mean, standard deviation, and credible intervals (95%) of the fixed effects and hyperparameters estimated by the MDAGAR models.

Table 4.2: Mean, standard deviation (SD), Lower Limit (LL) and Upper Limit (UL) of 95% credible intervals for the MDAGAR models with $\alpha = 0.4$, 0.6 and 0.9,

Response	Parameter	True value	Mean	SD	LL (95%)	UL (95%)
$\alpha = \mathbf{0.4}$						
y_1	β_{01}	2	2.001	0.09	1.821	2.177
	β_{11}	0.6	0.603	0.045	0.515	0.692
	τ_1	2	2.31	0.425	1.57	3.233
y_2	β_{02}	2.5	2.517	0.122	2.277	2.755
	β_{12}	-0.1	-0.095	0.053	-0.198	0.009
	τ_2	1	1.151	0.192	0.809	1.562
y_3	β_{03}	3	2.987	0.098	2.793	3.18
	β_{13}	0.2	0.176	0.043	0.092	0.26
	τ_3	1.5	1.766	0.287	1.256	2.381
	ρ_{12}	0.8	0.684	0.058	0.559	0.787
	ρ_{13}	0.6	0.407	0.08	0.242	0.554
	ρ_{23}	0.4	0.318	0.075	0.166	0.461
	α	0.4	0.389	0.068	0.267	0.532
$\alpha = \mathbf{0.6}$						
y_1	β_{01}	2	2	0.104	1.794	2.204
	β_{11}	0.6	0.636	0.039	0.561	0.712
	τ_1	2	1.978	0.511	1.139	3.132
y_2	β_{02}	2.5	2.513	0.139	2.239	2.788
	β_{12}	-0.1	-0.091	0.041	-0.171	-0.01
	τ_2	1	1.029	0.248	0.615	1.583
y_3	β_{03}	3	2.98	0.122	2.74	3.222
	β_{13}	0.2	0.179	0.035	0.109	0.248
	τ_3	1.5	1.333	0.315	0.805	2.037
	ρ_{12}	0.8	0.706	0.066	0.562	0.818
	ρ_{13}	0.6	0.449	0.084	0.272	0.6
	ρ_{23}	0.4	0.306	0.083	0.138	0.461
	α	0.6	0.651	0.067	0.515	0.777
$\alpha = \mathbf{0.9}$						
y_1	β_{01}	2	2.028	0.131	1.765	2.293
	β_{11}	0.6	0.585	0.029	0.528	0.643
	τ_1	2	2.002	0.854	0.75	4.044
y_2	β_{02}	2.5	2.546	0.197	2.152	2.945
	β_{12}	-0.1	-0.077	0.027	-0.13	-0.023
	τ_2	1	0.859	0.358	0.324	1.706
y_3	β_{03}	3	2.915	0.162	2.588	3.241
	β_{13}	0.2	0.174	0.022	0.131	0.218
	τ_3	1.5	1.286	0.547	0.476	2.584
	ρ_{12}	0.8	0.588	0.123	0.314	0.792
	ρ_{13}	0.6	0.351	0.132	0.073	0.588
	ρ_{23}	0.4	0.21	0.118	-0.026	0.434
	α	0.9	0.915	0.035	0.835	0.969

From Table 4.2, when α =0.4, the model does a good job estimating the parameters, but it is more difficult to estimate ρ_{12} and ρ_{13} which credible intervals do not include the true values. This can also be observed in the posterior marginal distributions showed in Figure 4.11. From Table 4.2, when α =0.6, the model does a good job estimating all the parameters. This can also be observed in the posterior marginal distributions showed in Figure 4.12. From Table 4.2, when α =0.9, the model does a good job estimating the parameters, but it is more difficult to estimate ρ_{12} and ρ_{13} which credible intervals do not include the true values. This can also be observed in the in the posterior marginal distributions showed in the third and forth rows of Figure 4.12.

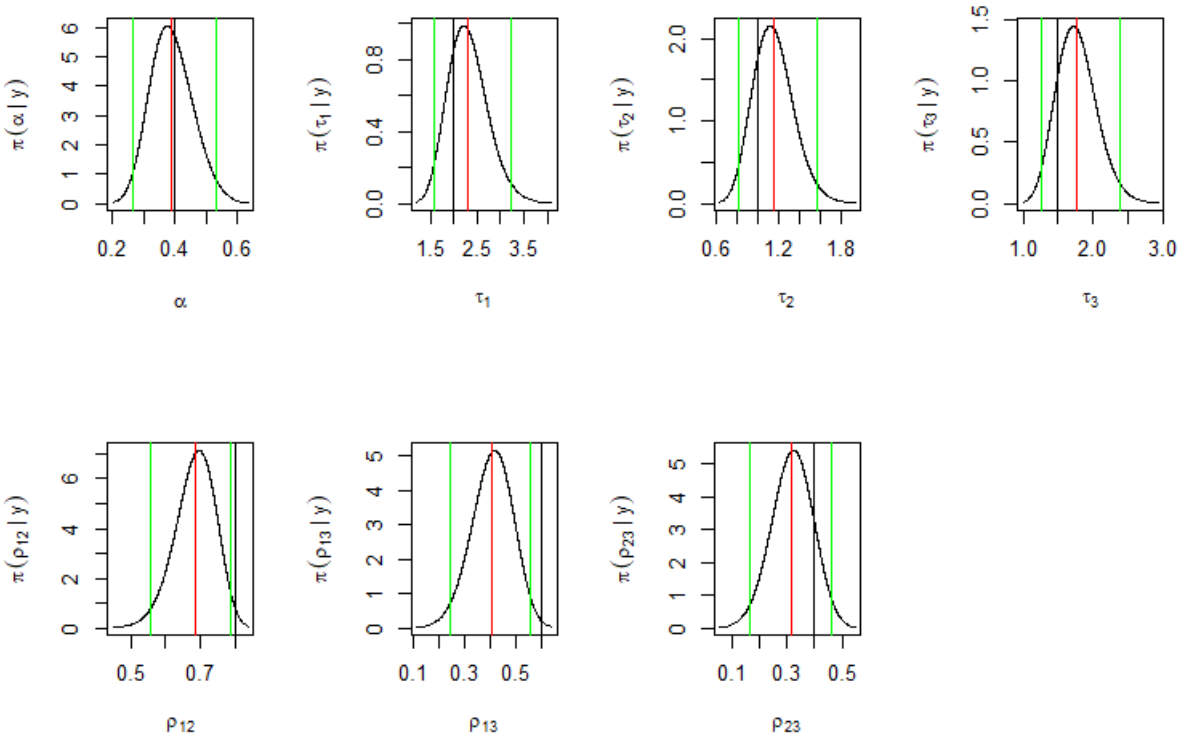


Figure 4.11: Marginal posterior density function of the hyperparameters estimated using the MDAGAR model for α = 0.1. The black vertical line indicates the true value of the hyperparameter. The red line indicates the mean of the estimated distribution. The green lines indicates the lower and upper limits of the 95% credible interval.

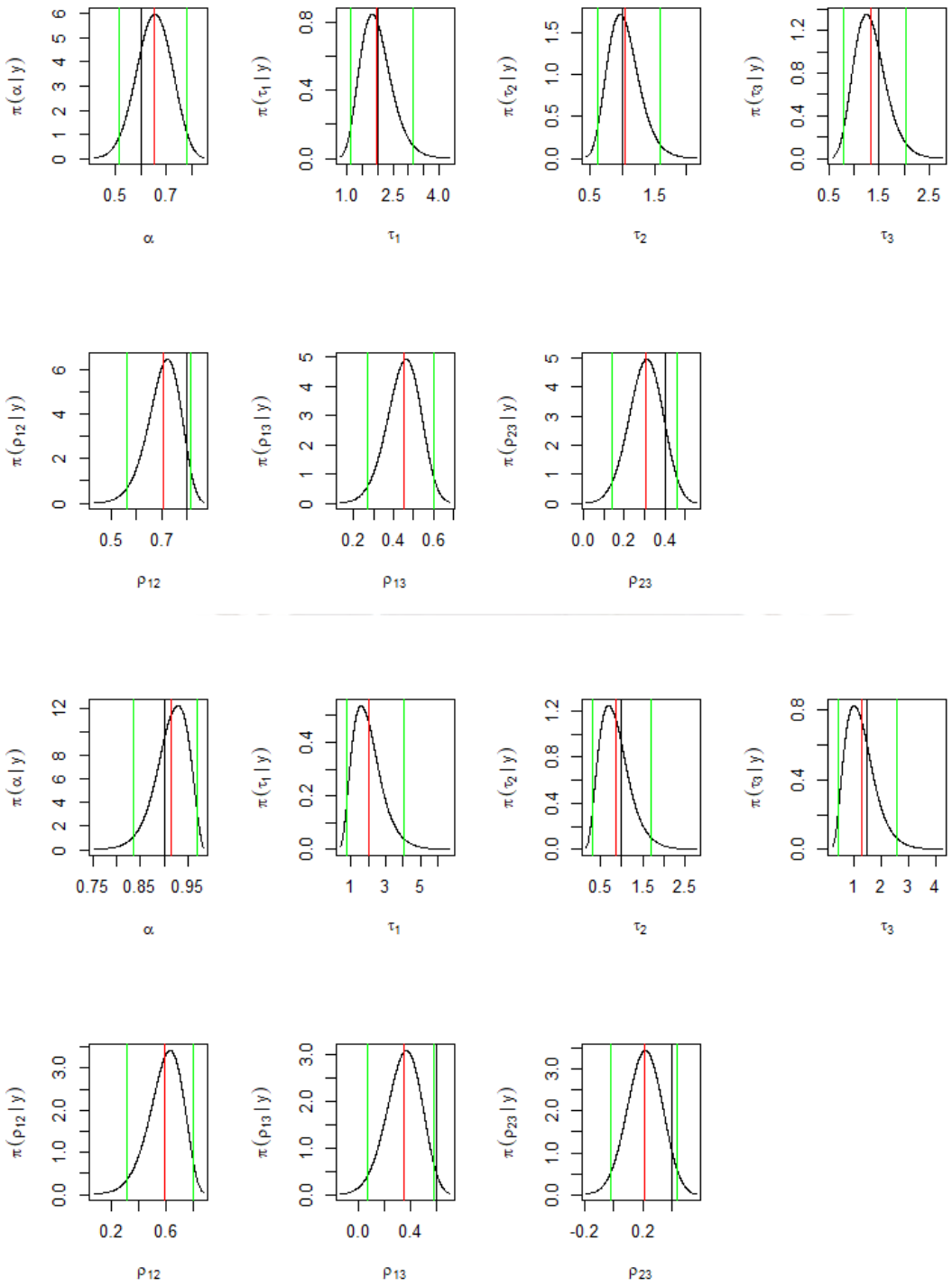


Figure 4.12: Marginal posterior density function of the hyperparameters estimated using the MDAGAR model. The black vertical line indicates the true value of the hyperparameter. The red line indicates the mean of the estimated distribution. The green lines indicates the lower and upper limits of the 95% credible interval. For α =0.6 (first and second row) and α =0.9 (third and fourth row).

The spatial random effects and response variables are also well estimated as can be seen in Figures 4.13 and 4.14, respectively. It can be observed that the estimates tend to be closer to the true value as α decreases for both the spatial random effects and the response variables. We can see that the response variables Y_2 and Y_3 are estimated better than the response variable Y_1 for any value of α.

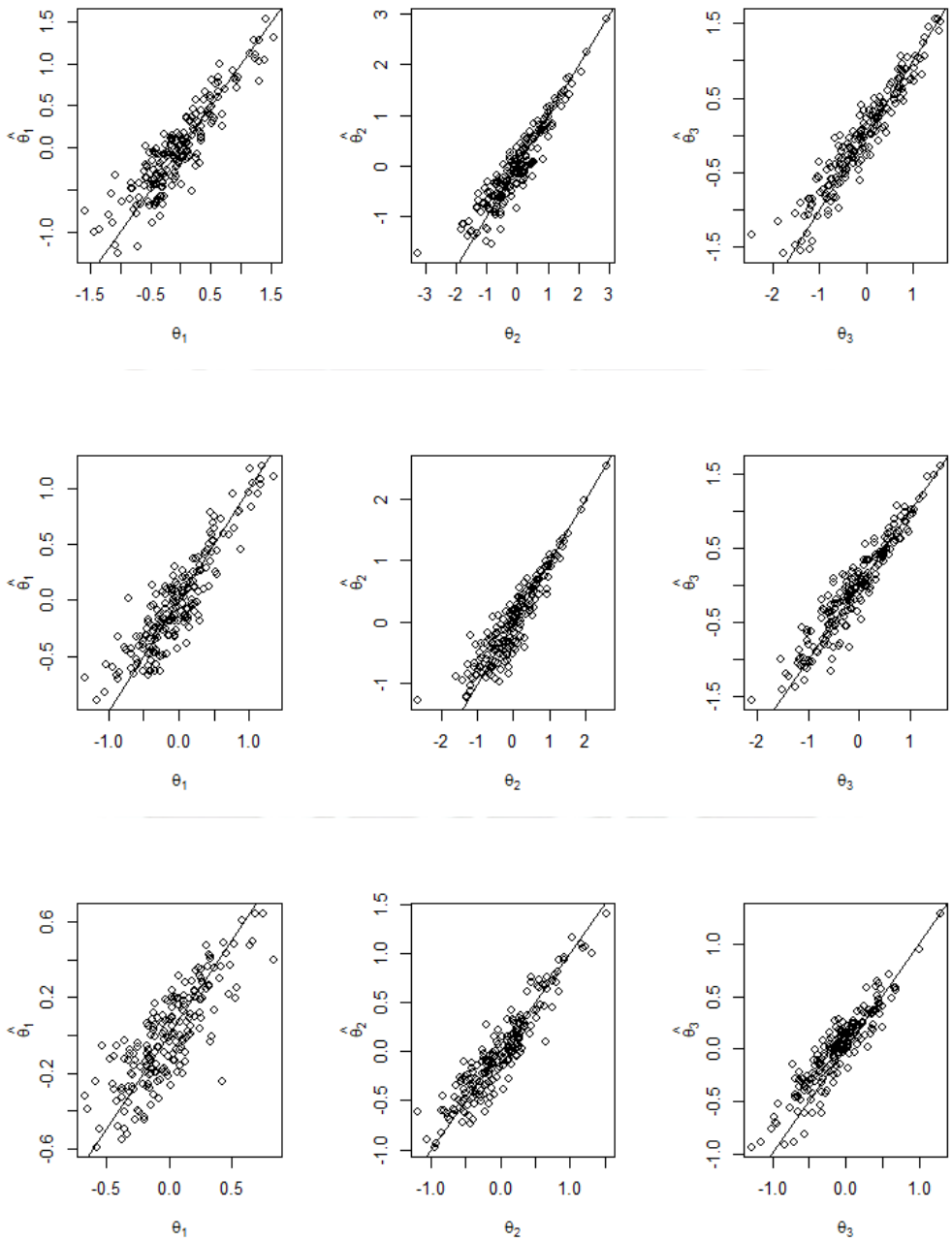


Figure 4.13: Plots of the spatial random effects corresponding to the three response variables against their estimations from the MDAGAR model. A line of slope equal to 1 is plotted for reference. For α =0.4 (first row), α =0.6 (second row) and α =0.9 (third row).

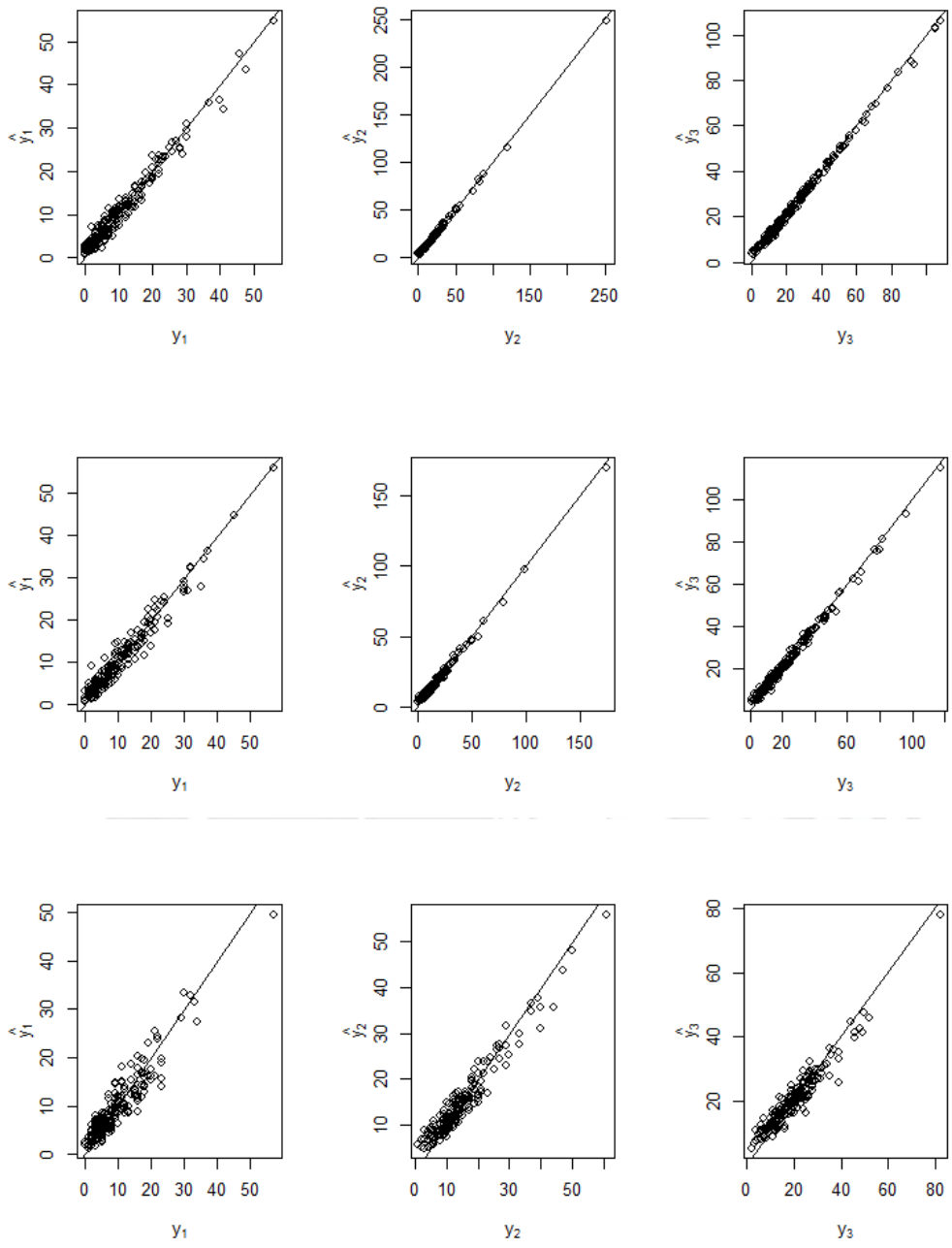


Figure 4.14: Plots of the three response variables simulated against their estimations from the MDAGAR model. A line of slope equal to 1 is plotted for reference. For α =0.4 (first row), α =0.6 (second row) and α =0.9 (third row).

In summary, the MDAGAR model does a good job estimating the parameters of the simulated model, performing better as α decreases. However, these results are only for one simulation and more simulations are needed to draw conclusions as we present in the next section.

4.3 Simulation 3: MCAR versus MDAGAR models

In this section, we compare the performance of both models under two scenarios: i) In Scenario I, we simulate data from MCAR model, but fit MCAR and MDAGAR models, ii) In Scenario II, we simulate data from MDAGAR model, but fit MCAR and MDAGAR models. In both scenarios, the data were simulated from MCAR and MDAGAR models for $\alpha = 0.2, 0.4, 0.6, 0.8, 0.9$. We also run 100 replicates for each scenario.

4.3.1 Scenario I

Figure 4.15 shows the boxplots of the posterior medians of the fixed effects estimated in the 100 replicates for each α. It can be observed that there is little difference between the estimations made by MCAR and MDAGAR with both models giving really good estimations.

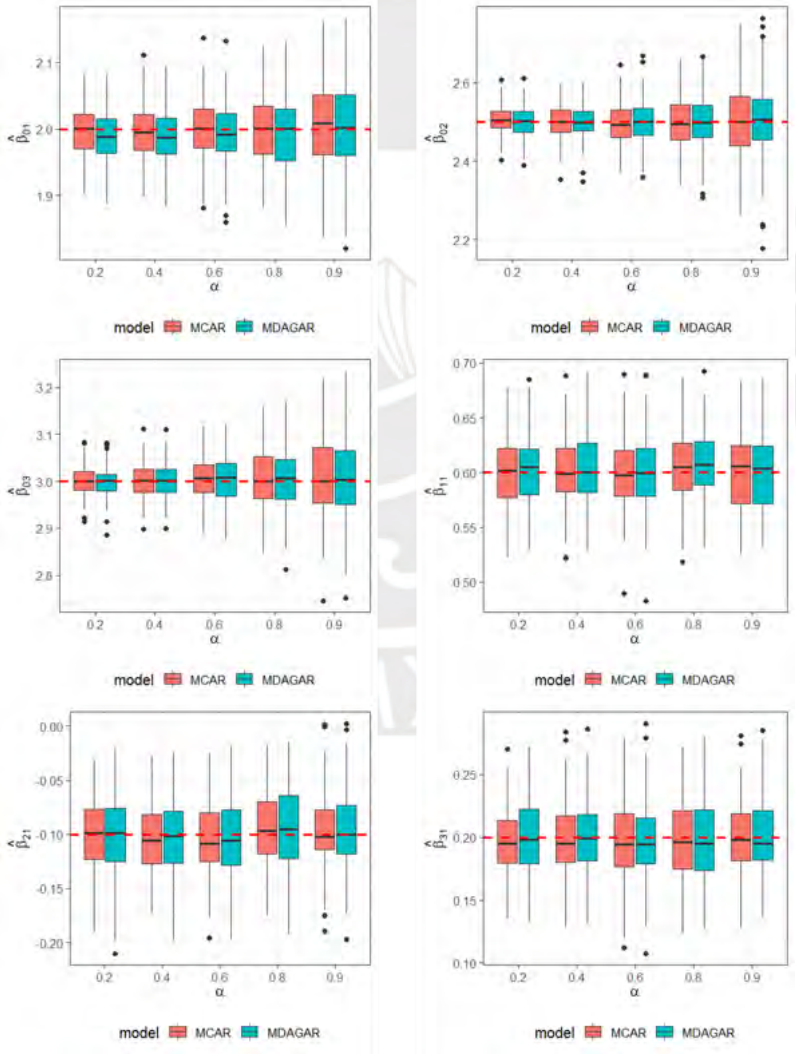


Figure 4.15: Scenario I. Boxplots of the medians of the estimated posterior marginal distributions of the fixed effects. The true value is represented by the red dashed line.

Figure 4.16 shows the coverage probabilities of the posterior medians of the fixed effects. The coverages were computed as the proportion of the estimations (of the 100) where 95% credible interval includes the true value. It can be observed that a large proportion of the 100 simulations estimated a credible interval that includes the true value, the proportion being almost close to one across all values of α in every fixed effect.

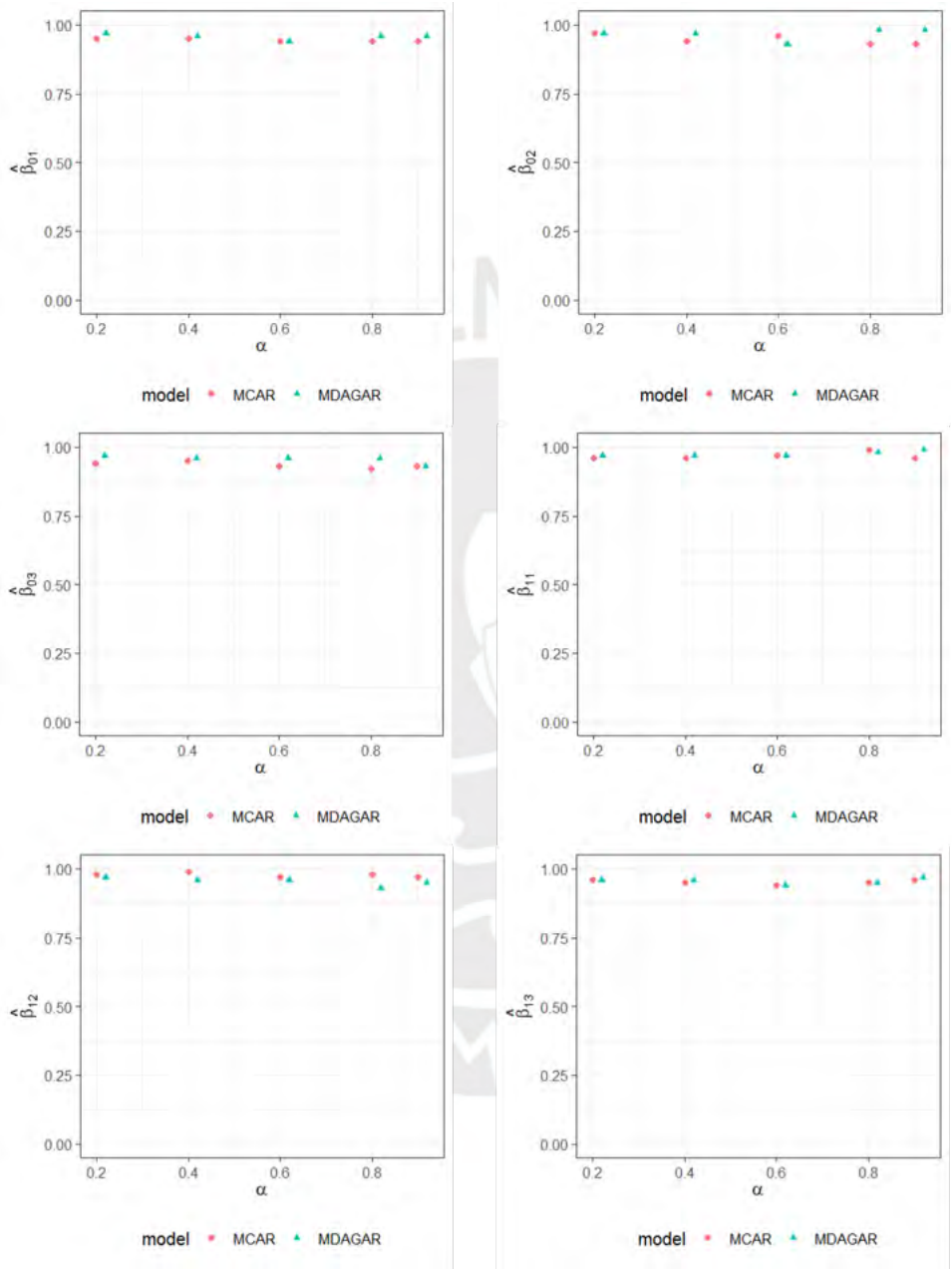


Figure 4.16: Scenario I. Coverage probabilities of the estimated posterior marginal distributions of the fixed effects.

Figure 4.17 shows the boxplots of the posterior medians of the hyperparameters estimated in the 100 replicates for the different values of α. As expected, it can be observed that the estimations of the median of α through the MCAR models are closer to the true value than

the MDAGAR estimations, however the MDAGAR models are not so far away from the true values. Regarding the hyperparameters of correlation between response variables, both models estimate better this parameter when it is lower. However, it should be noticed that the performance of MDAGAR model is slightly better for higher values of α. This result is interesting, because although the data were simulated from MCAR model, the same MCAR model is not able to recover very well these correlation parameters between response variables, specially when the correlation between response variables is higher.

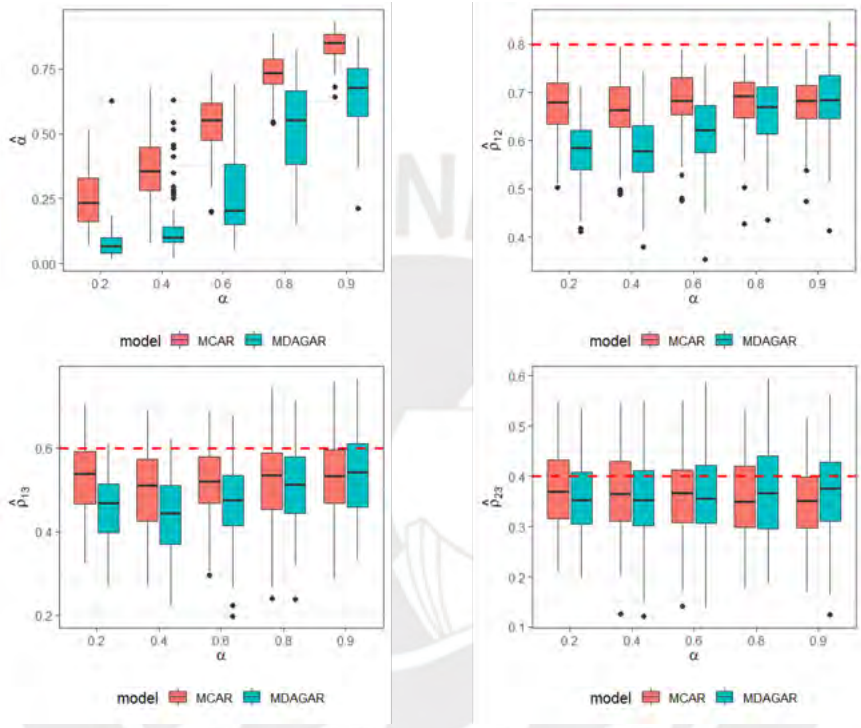


Figure 4.17: Scenario I. Boxplots of the medians of the estimated posterior marginal distributions of the hyperparameters of spatial autocorrelation and correlation between variables. The true value is represented by the red dashed line.

Regarding the hyperparameters of precision, Figure 4.18 shows that the MCAR model gives better estimates than the MDAGAR model but this difference decreases for larger values of α. Also as expected, the 100 estimates of the MCAR model have less variability for all values of α.

Figure 4.19 shows the coverage probabilities of the hyperparameters of spatial autocorrelation and correlation between variables. This Figure shows that the coverage probabilities of the MCAR model are very close to one for α and for the correlation between the second and third variable ρ_{13}, while its performance decreases for the other correlation hyperparameters with higher values. Regarding the MDAGAR model, the coverage probabilities are close to one for ρ_{13}. However, as expected in general, for the other hyperparameters the coverage probabilities are far from one, but they get closer to one as α increases.

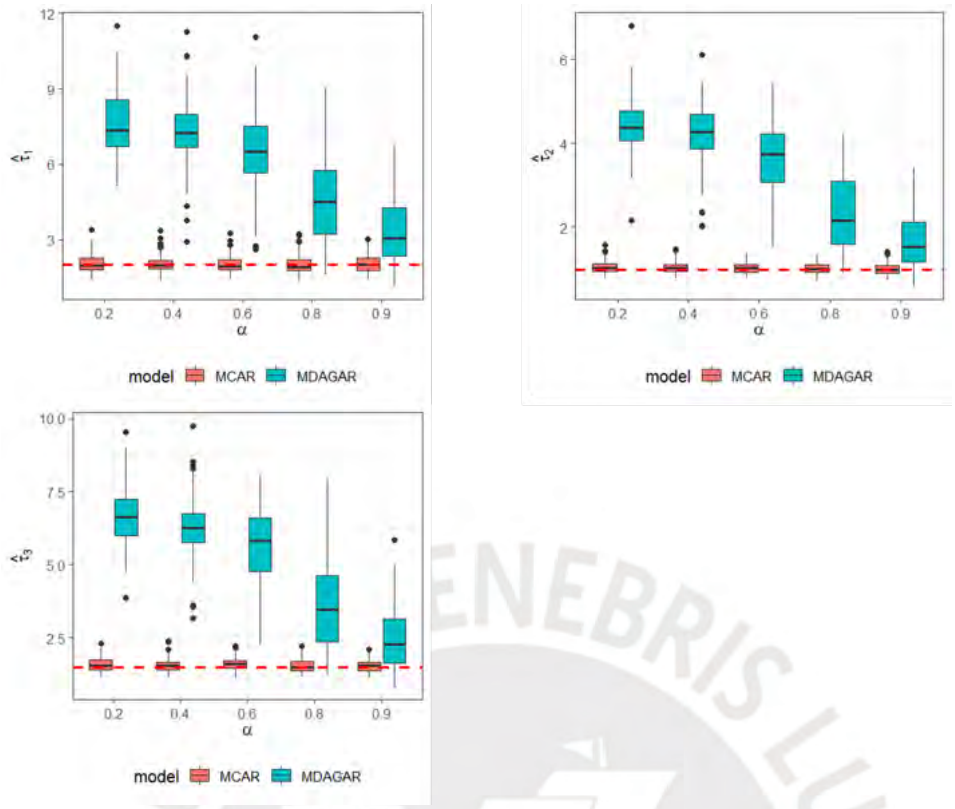


Figure 4.18: Scenario I. Boxplots of the medians of the estimated posterior marginal distributions of the precision hyperparameters. The true value is represented by the red dashed line.

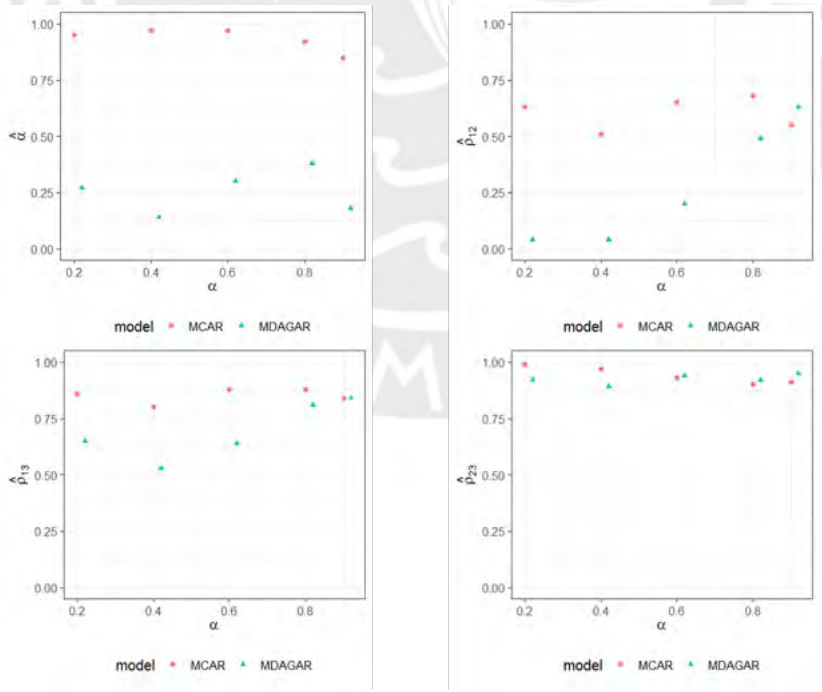


Figure 4.19: Scenario I. Coverage probabilities of the estimated posterior marginal distributions of the hyperparameters of spatial autocorrelation and correlation between variables.

Figure 4.20 shows the coverage probabilities of the precision hyperparameters which are also very close to one for the MCAR model while for the MDAGAR model the coverage probabilities increase as α increases.

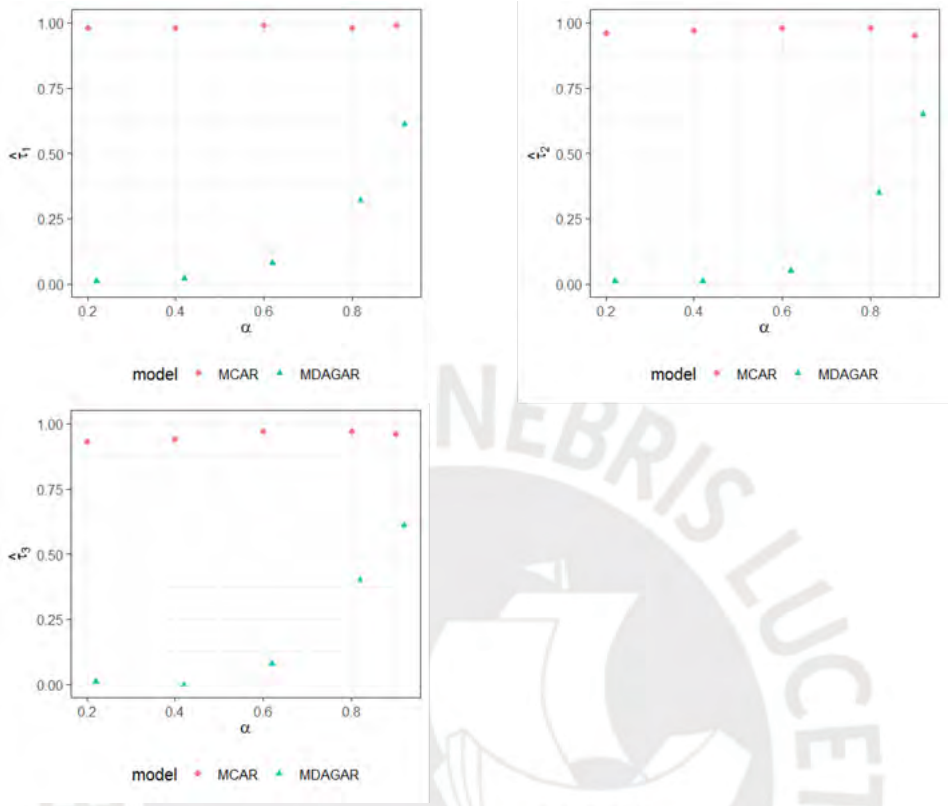


Figure 4.20: Scenario I. Coverage probabilities of the estimated posterior marginal distributions of the precision hyperparameters.

4.3.2 Scenario II

Figure 4.21 shows the boxplots of the posterior medians of the fixed effects estimated in the 100 replicates for each α. It can be observed that there is little difference between the estimations made by MCAR and MDAGAR, both models giving really good estimations. While, Figure 4.22 shows the coverage probabilities of the fixed effects. It can be observed that a large proportion of the 100 simulations estimated a credible interval that includes the true value, the proportion being almost close to one across all values of α in every fixed effect.

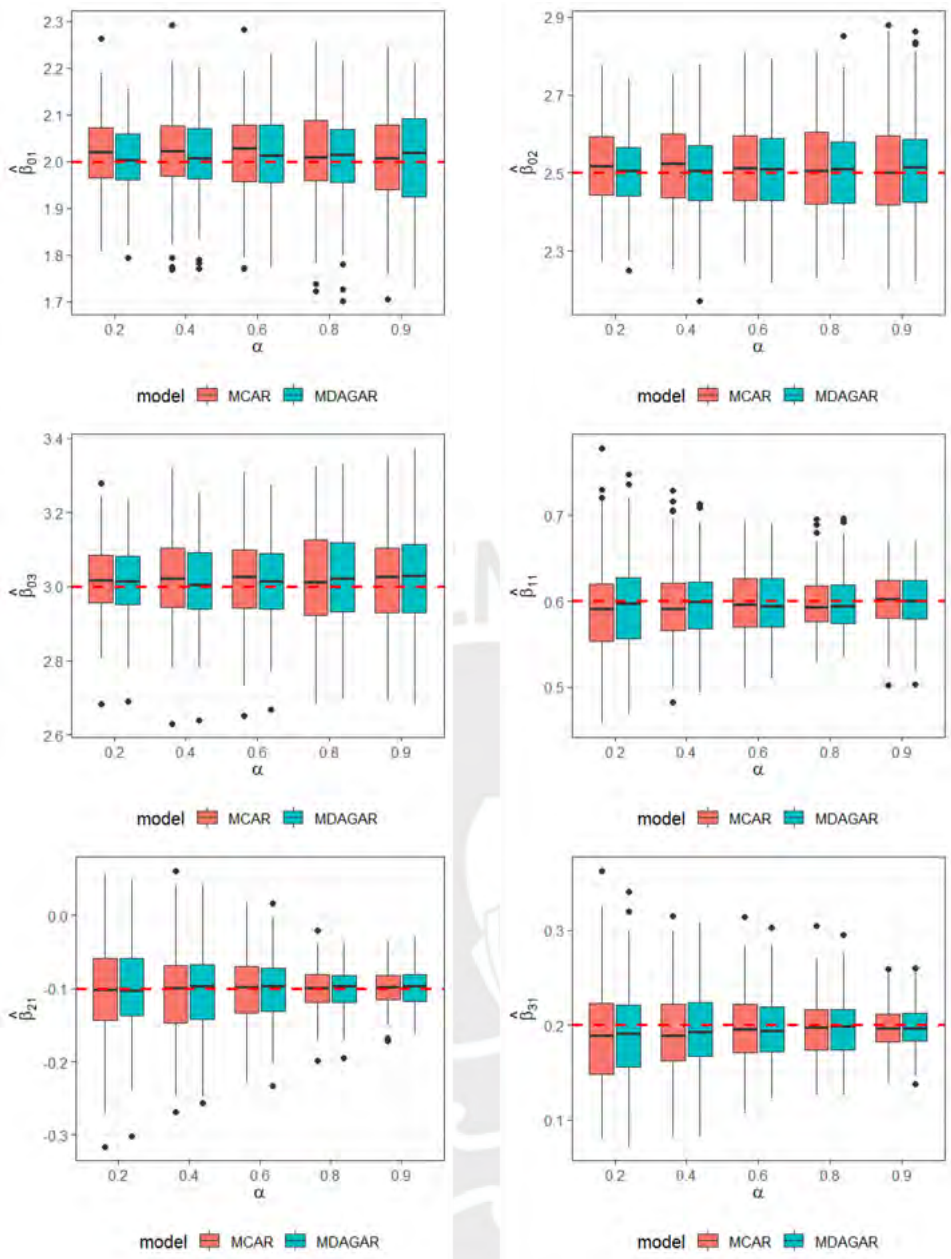


Figure 4.21: Scenario II. Boxplots of the medians of the estimated posterior marginal distributions of the fixed effects. The true value is represented by the red dashed line.

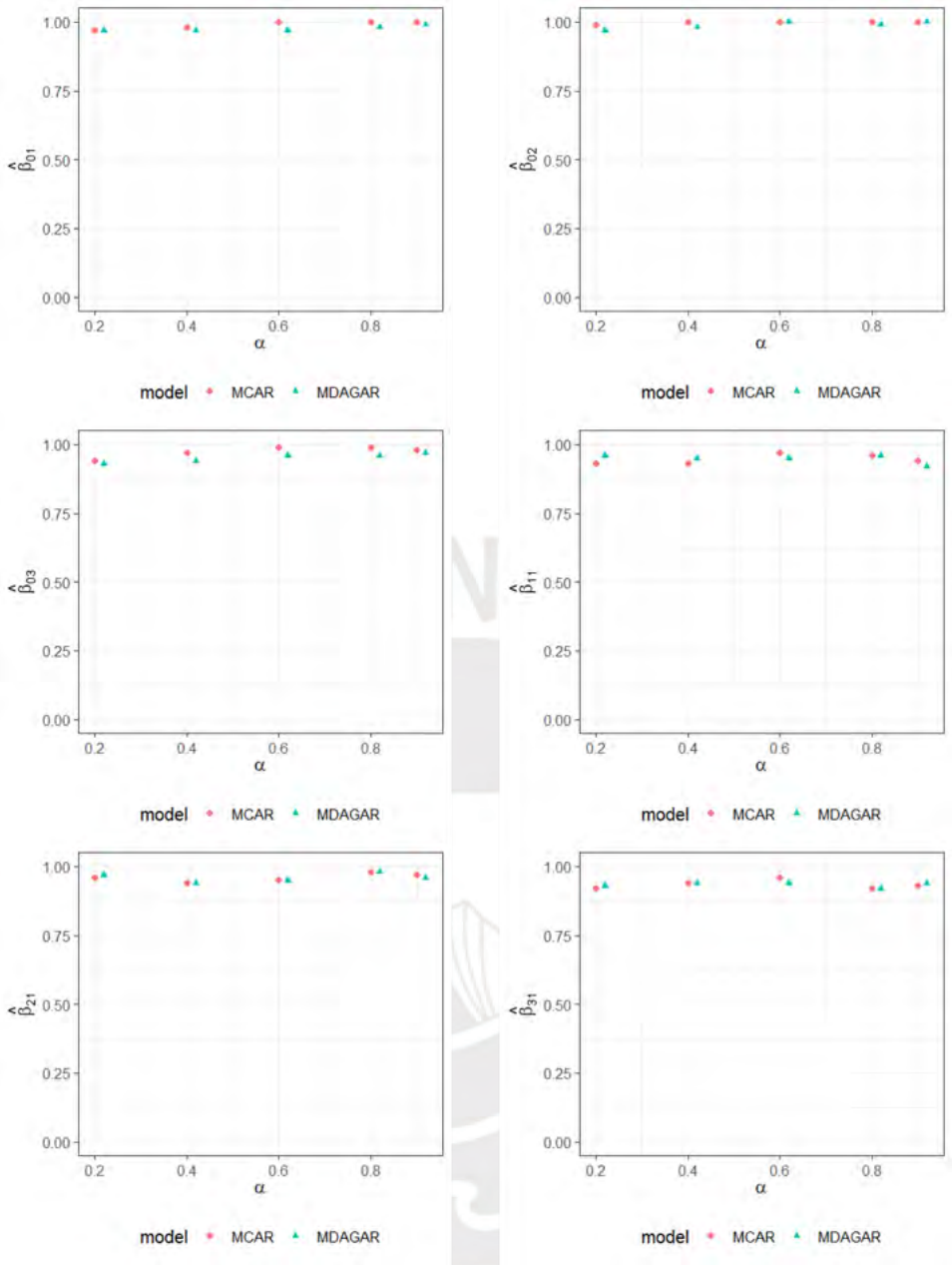


Figure 4.22: Scenario II. Coverage probabilities of the estimated posterior marginal distributions of the fixed effects.

Figure 4.23 shows boxplots of the posterior medians of the hyperparameters estimated in the 100 replicates for each α. Regarding α, the MDAGAR model posterior estimates are really close to the true values, while the MCAR model tends to overestimate this value. This overestimation is stronger for smaller values of α. Regarding the hyperparameters of correlation between response variables, both models tend to improve their estimations as α decreases, but in general, the MDAGAR model is closer to the true values. Regarding the hyperparameters of precision, as expected Figure 4.24 shows that the MDAGAR model gives better posterior estimates than the MCAR model. The MCAR model tends to underestimate

the true precision for small values of α and it tends to overestimate the true precision for big values of α.

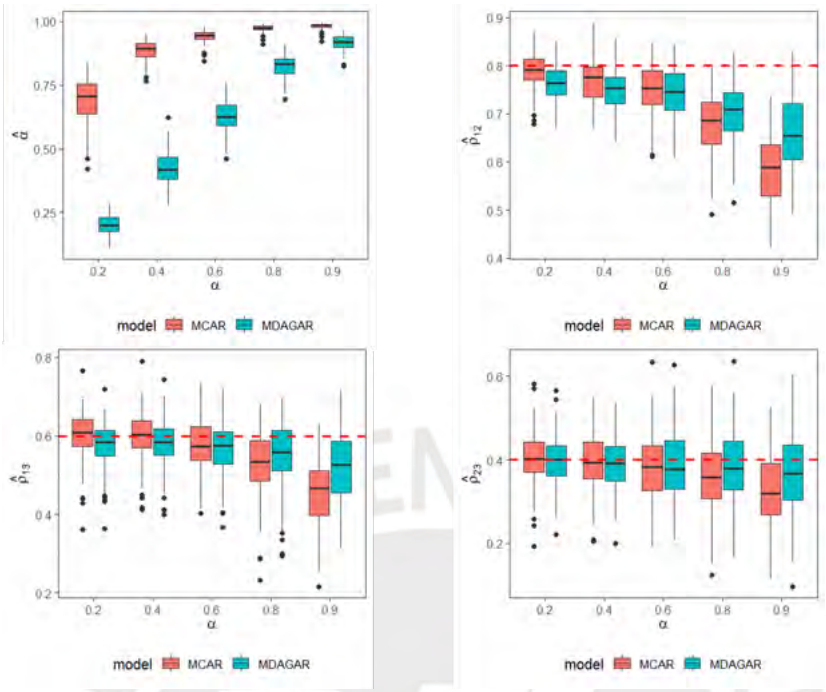


Figure 4.23: Scenario II. Boxplots of the medians of the estimated posterior marginal distributions of the hyperparameters of spatial autocorrelation and correlation between variables. The true value is represented by the red dashed line.

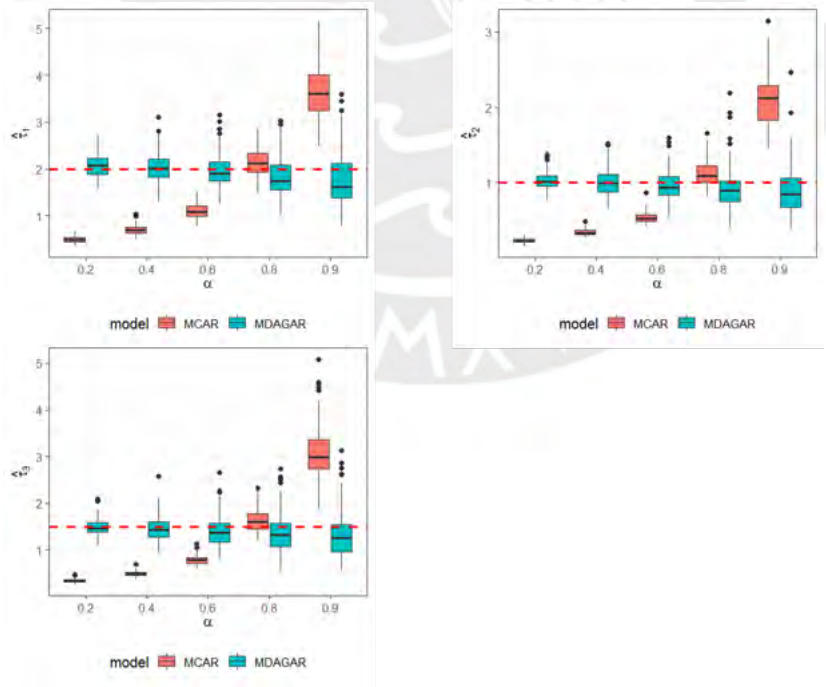


Figure 4.24: Scenario II. Boxplots of the medians of the estimated posterior marginal distributions of the precision hyperparameters. The true value is represented by the red dashed line.

Figure 4.25 shows the coverage probabilities of the hyperparameters. It can be observed that for the MDAGAR model the coverage probabilities are very close to one for all the parameters, except for the correlation between the first and second variable, which performance decreases as α increases. Regarding the MCAR model, the coverage probabilities are close to one for the correlation between the first and second, and second and third response variables. Nevertheless, for the other hyperparameters, in general the coverage probabilities are very far from one.

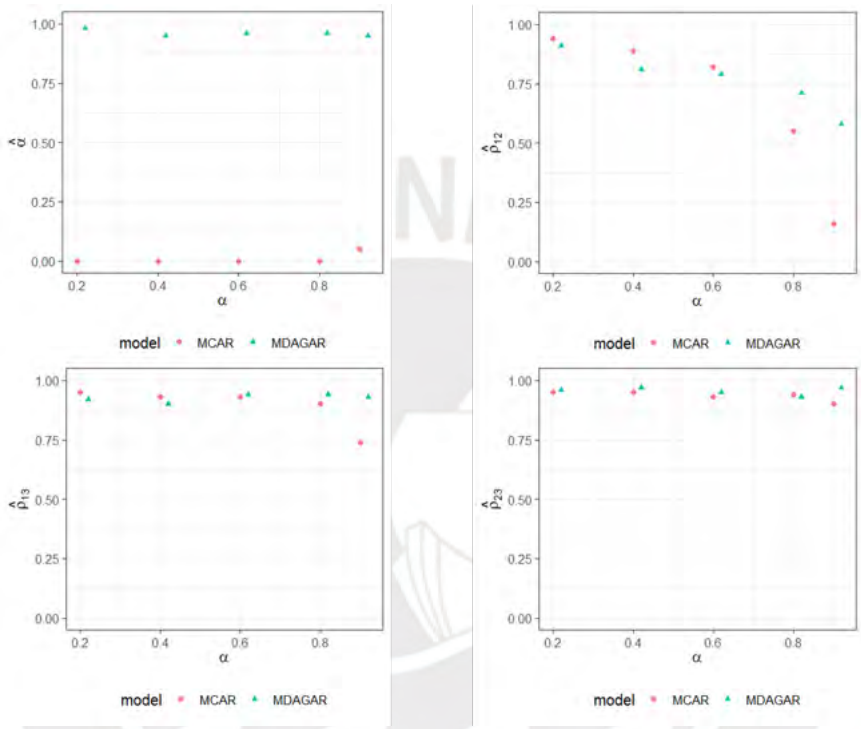


Figure 4.25: Scenario II. Coverage probabilities of the estimated posterior marginal distributions of the hyperparameters of spatial autocorrelation and correlation between variables.

Figure 4.26 shows the coverage probabilities of the precision hyperparameters which are also very close to one for the MDAGAR model while for the MCAR model the coverage probabilities are very far from one.

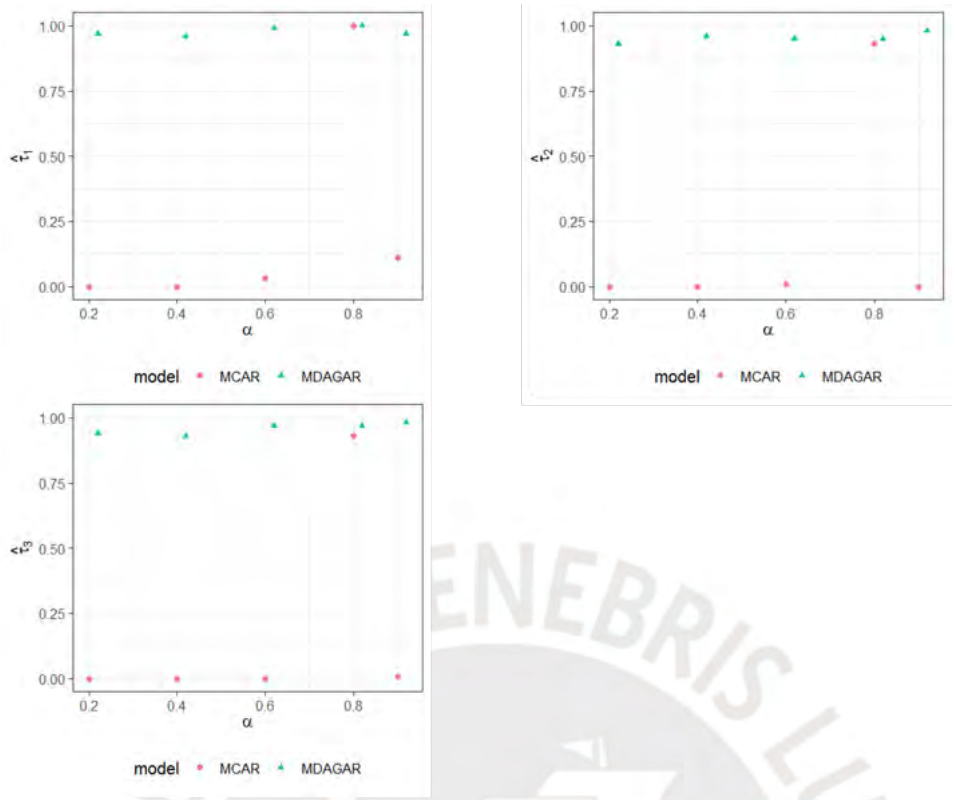


Figure 4.26: Scenario II. Coverage probabilities of the estimated posterior marginal distributions of the precision hyperparameters.

4.4 Simulation 4: Simulation from an Exponential Gaussian Process

In this section, we simulate data from an exponential Gaussian process and we compare the performance of both models at the same time fitting the MCAR and MDAGAR models. The data were simulated for $\alpha = 0.2, 0.4, 0.6, 0.8, 0.9$. We run 100 replicates for each scenario.

Figure 4.27 shows the boxplots of the posterior medians of the fixed effects estimated in the 100 replicates for each α. It can be observed that there is little difference between the estimations made by MCAR and MDAGAR with both models giving really good estimations, and more dispersion as α increases.

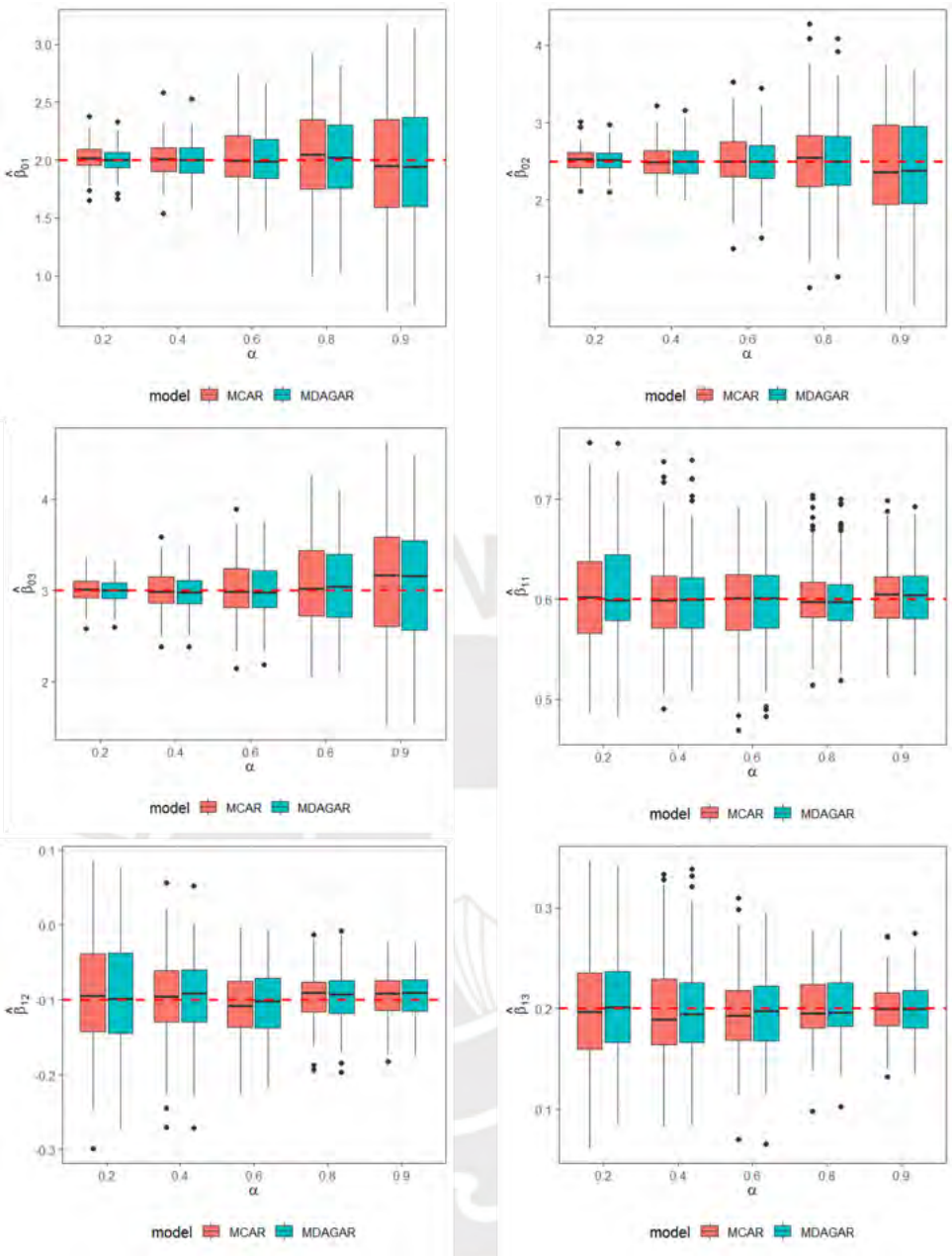


Figure 4.27: Boxplots of the medians of the estimated posterior marginal distributions of the fixed effects. The true value is represented by the red dashed line.

Figure 4.28 shows the coverage probabilities of the posterior medians of the fixed effects. Two different patterns can be observed in this figure. For the intercepts of the linear predictor, the MCAR model has higher coverage probabilities than the MDAGAR model in every scenario, and the coverage probabilities of the MDAGAR model is higher for small values of α. For the slopes of the linear predictor, the coverage probabilities are very close to one in both models for every α.

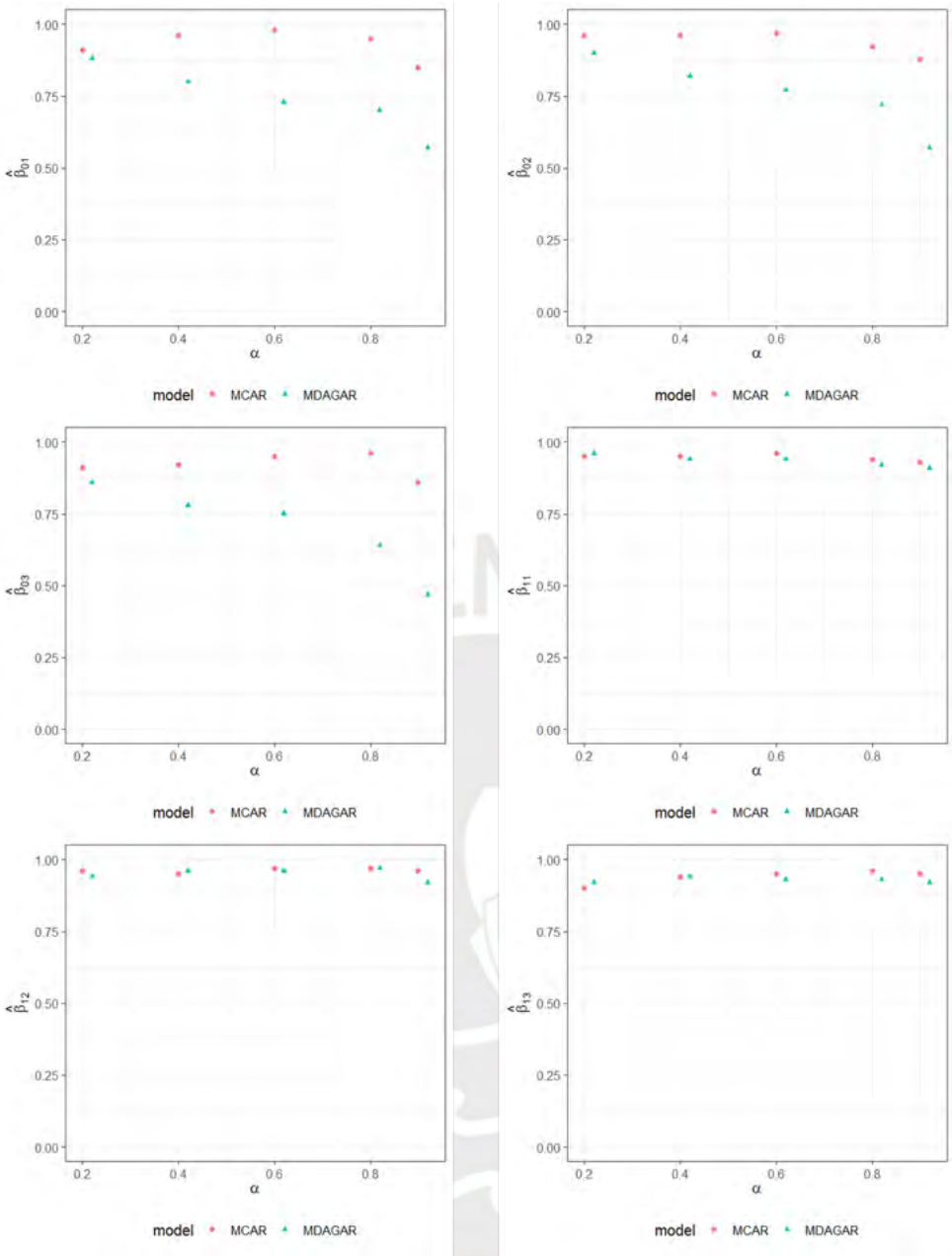


Figure 4.28: Coverage probabilities of the estimated posterior marginal distributions of the fixed effects.

Figure 4.29 shows the boxplots of the posterior medians of the hyperparameters estimated in the 100 replicates for the different values of α. It can be observed that the estimations of the median of α through the MDAGAR models are closer to the true value than the MCAR estimations, which overestimates the true value of α, specially as α increases. Regarding the hyperparameters of correlation between response variables, both models perform well with a little underestimation for values of α higher than 0.6. However, it should be noticed that the performance of MDAGAR model is slightly better for higher values of α.

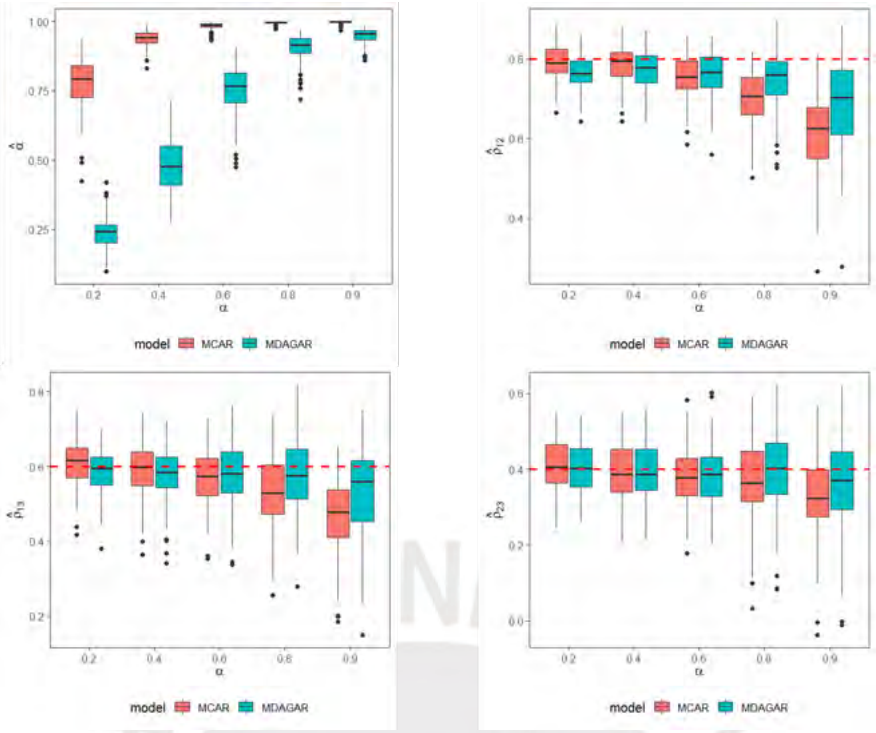


Figure 4.29: Boxplots of the medians of the estimated posterior marginal distributions of the hyperparameters. The true value is represented by the red dashed line.

Figure 4.30 shows the coverage probabilities of the hyperparameters. It can be observed that the coverage probabilities of the MCAR model are zero or they are very close to zero for α, while the MDAGAR model tends to recover the true value around 75% of the time for α less than 0.5, and around 40% of the time for α higher than 0.5. Regarding the correlation between the second and third variable, both models have very high coverage probabilities, being slightly better for MDAGAR models. For the other two hyperparameters of correlation, both models' performance tend to decrease as α increases, with the MDAGAR model having the best performance.

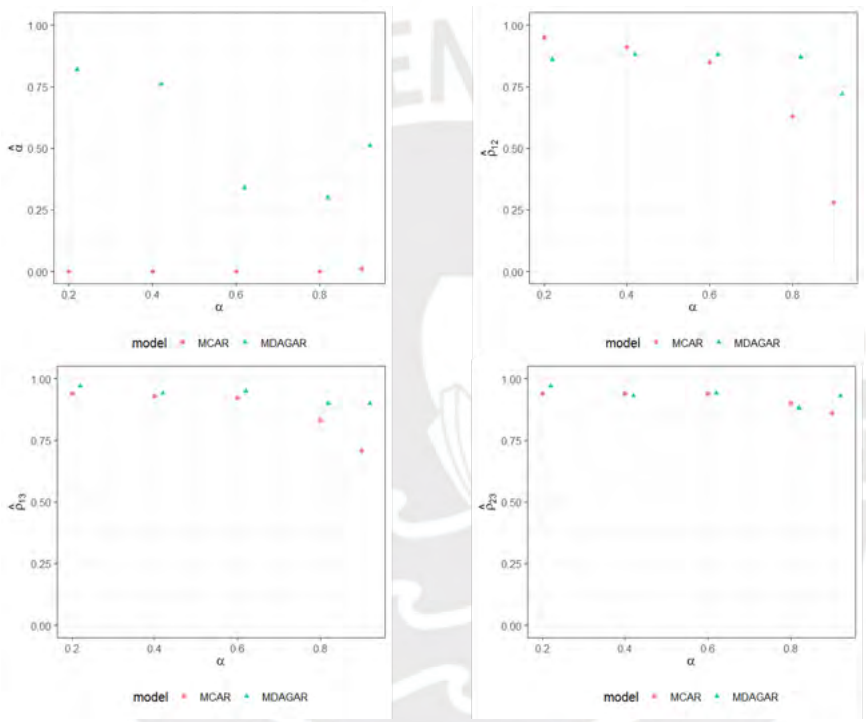


Figure 4.30: Coverage probabilities of the estimated posterior marginal distributions of the hyperparameters.

Chapter 5

Applications

In this chapter, the MDAGAR model is applied to fit three datasets, and we compared the results and performance of this model against the MCAR model.

5.1 Application 1: Mortality in Comunidad Valenciana

This dataset contains the number of deaths caused by cirrhosis, lung and oral cancer as well as the expected number of cases in 540 districts of Comunidad Valenciana (Spain). The dataset was analyzed in Palmí-Perales et al. (2021) where more description of the data is available.

We assume that Y_{id} represents the observed number of deaths from cirrhosis ($d = 1$), lung cancer ($d = 2$) and oral cancer ($d = 3$) in municipality i, for $i = 1, 2, \ldots, 540$. The relative risk (R_{id}) is defined as the number of deaths (Y_{id}) divided by the expected number of deaths (E_{id}). Figure 5.1 shows the relative risk for each disease in the 540 municipalities of Comunidad Valenciana and Figure 5.2 shows the histogram of the relative risk of each disease.

We assume a multivariate Poisson distribution for Y_{id} such that:

$$Y_{id} \sim Poisson(\mu_{id} = E_{id}R_{id}); \quad i = 1, 2, \ldots, n = 540; \quad d = 1, 2, 3,$$

where μ_{id} represents the average of deaths of each disease d in the i-th municipality. This mean μ_{id} is the modeled throughout:

$$\log(\mu_{id}) = \log(E_{id}) + \beta_{0d} + \theta_{id},$$

where E_{id} is the offset, β_{0d} is the intercept of the d-th disease and θ_{id} represents the spatial

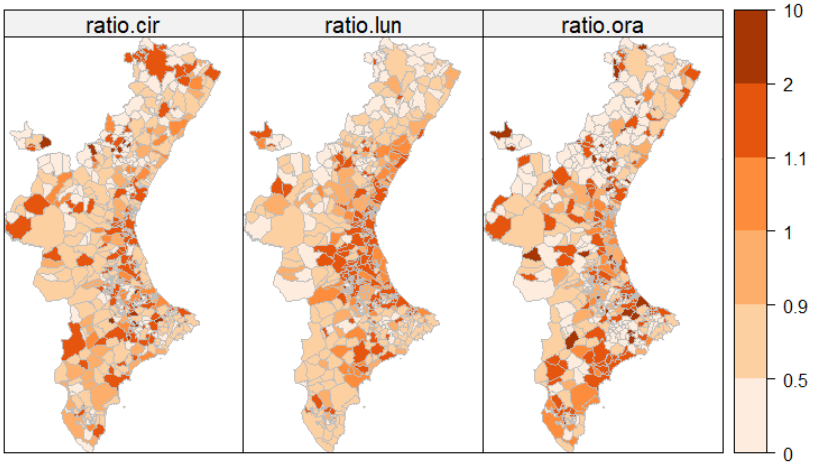


Figure 5.1: Maps of the relative risk for each disease in Comunidad Valenciana.

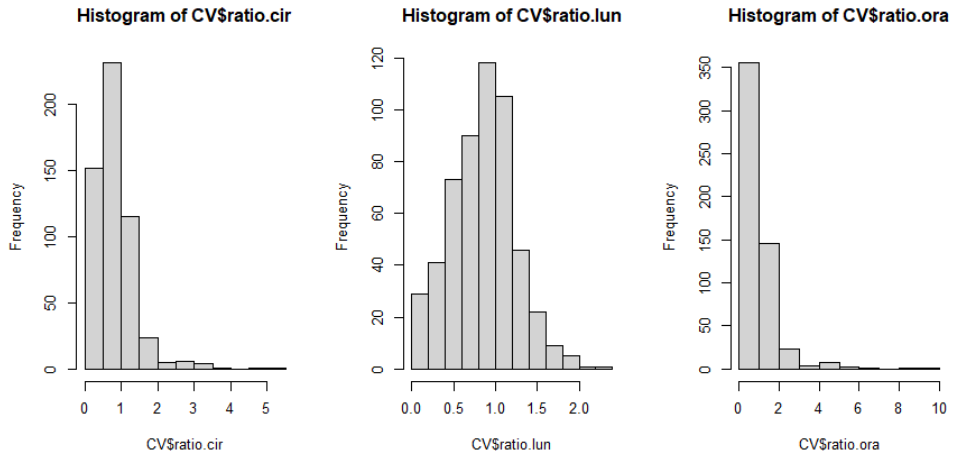


Figure 5.2: Histograms of the relative risk of each disease in Comunidad Valenciana.

random effect. We assume that these spatial random effects are defined from a MCAR and MDAGAR model, then we fit the MCAR and MDAGAR models through INLA.

Model selection criteria such as the Watanabe-Akaike Information Criterion (WAIC), Logarithm of the Pseudo Marginal Likelihood (LPML), and the mean squared error (MSE) were calculated. Table 5.1 shows the mean and 95 % credible intervals for the parameters of each model, as well as the model selection criteria. The MCAR model shows better goodness of fit than the MDAGAR model as can be seen by the lower WAIC, and MSE of the MCAR model. However, the MDAGAR model presents a better fit according to its higher LPML value.

Table 5.1: Mean and credible intervals (95%) for the estimated parameters of MCAR and MDAGAR models, as well as the model selection metrics.

Response	Parameter	MCAR	MDAGAR
y_1	β_{01}	-0.393 (-0.710 , -0.080)	-0.527 (-0.806 , -0.270)
	τ_1	2.352 (1.756 , 3.085)	0.760 (0.213, 1.718)
y_2	β_{02}	-0.223 (-0.423 , -0.025)	-0.298 (-0.462 , -0.144)
	τ_2	5.769 (4.481 , 7.324)	2.033 (0.639 , 4.332)
y_3	β_{03}	-0.571 (-0.954 , -0.199)	-0.721 (-1.052 , -0.417)
	τ_3	1.644 (1.178 , 2.233)	0.544 (0.159 , 1.204)
	ρ_{12}	0.608 (0.472 , 0.724)	0.683 (0.542 , 0.797)
	ρ_{13}	0.770 (0.652 , 0.859)	0.836 (0.720 , 0.916)
	ρ_{23}	0.781 (0.681 , 0.857)	0.844 (0.741 , 0.918)
	α	0.991 (0.977 , 0.998)	0.934 (0.857 , 0.981)
WAIC		**7528.031**	7594.694
LPML		-4054.99	**-3920.713**
MSE		**0.467**	0.477
Time (sec.)		11.1	177

In Table 5.1 we also observe that the mean posterior of the spatial autocorrelation parameter α through the MCAR model is higher than with the MDAGAR model, being 0.991 and 0.934, respectively. Moreover, the credible interval for α is smaller for the MCAR model, with values very close to one, while the MDAGAR models estimate a credible interval for α between 0.857 and 0.981. We also observe that the posterior mean estimates of the precision parameters are also very different between the MCAR and MDAGAR models, being higher in the MCAR model, which indicates that less spatial variability is captured by this model. The rest of parameter estimates are similar for both models.

The posterior marginal distributions from MCAR and MDAGAR models are showed together in Figure 5.3. As can be seen, the marginal distribution of precision hyperparameters from both models are very different, while the hyperparameters of correlation between diseases are more similar. Regarding the spatial autocorrelation parameter α, we also observe that the MCAR model estimation present a smaller variance than the MDAGAR model.

The posterior marginal distributions from MCAR and MDAGAR models are also showed in Figure 5.4 and Figure 5.5, respectively, where the mean is represented by the red vertical line and 95 % credible interval is represented by the green vertical lines. In Figure 5.4 we observe that the MCAR model estimated a very high mean α of 0.991 with credible interval [0.977, 0.998], suggesting a spatial autocorrelation almost equal to 1, while the MDAGAR model gives a slightly smaller estimate of the mean of $\alpha = 0.934$ with a large credible interval of [0.857, 0.981], giving chance of a smaller spatial autocorrelation.

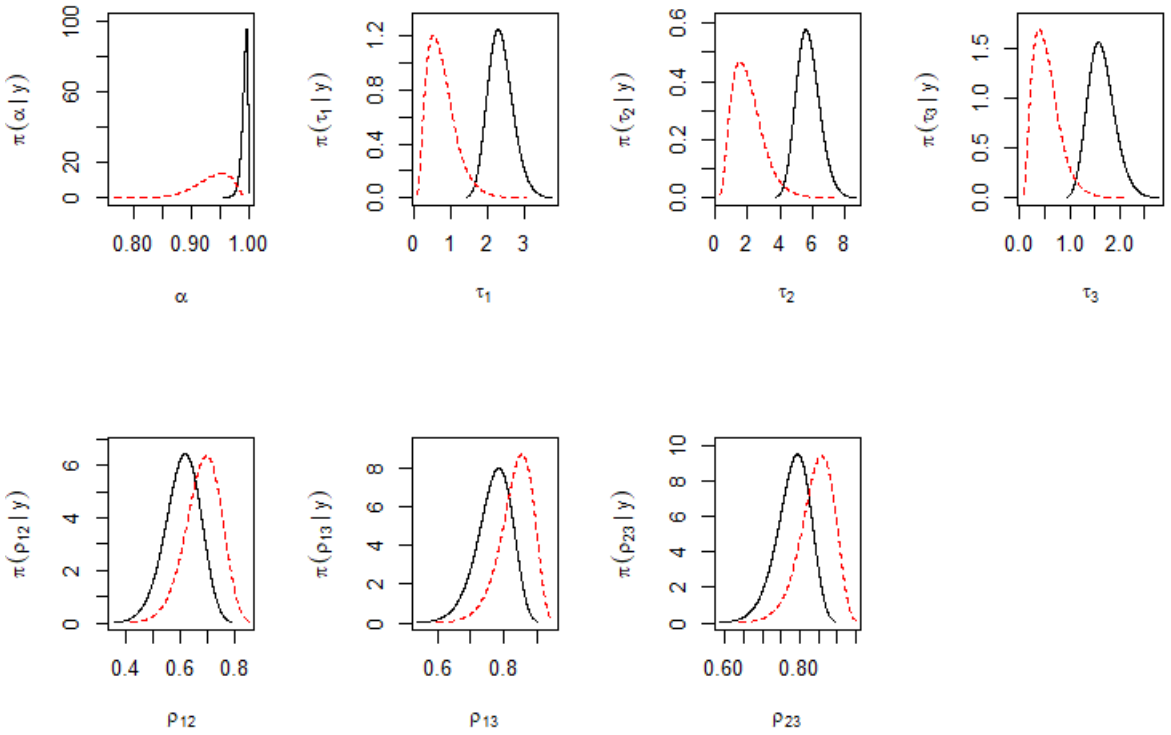


Figure 5.3: Probability density function of the hyperparameters estimated by MCAR (black) and MDAGAR (red) models.

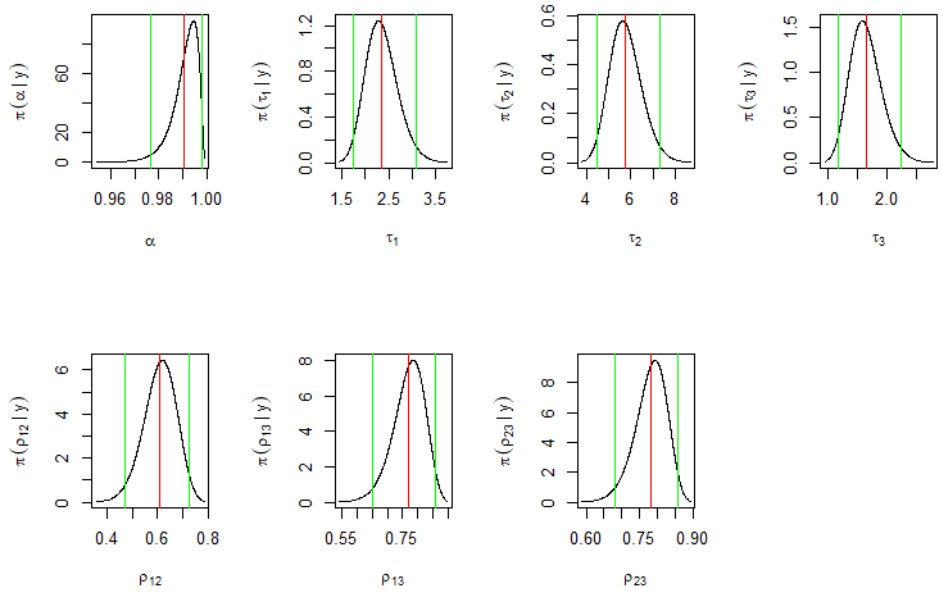


Figure 5.4: Probability density function (black), mean (red) and 95 % credible interval (green) of the hyperparameters estimated by MCAR model.

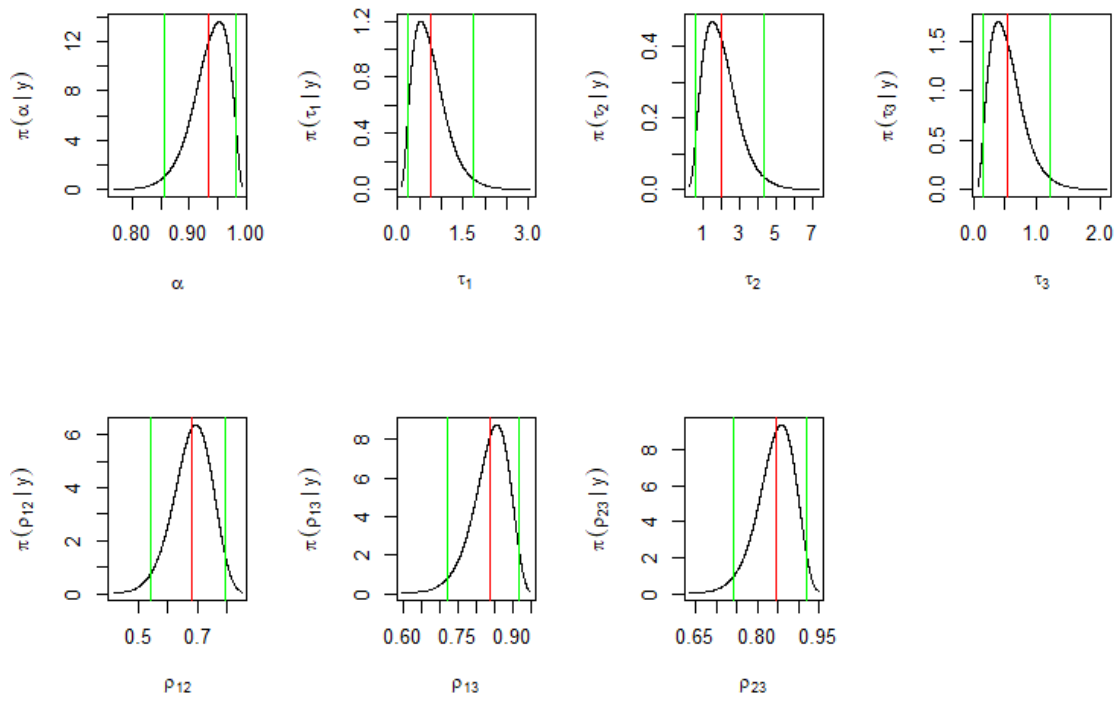


Figure 5.5: Probability density function (black), mean (red) and 95 % credible interval (green) of the hyperparameters estimated by MDAGAR model.

Figure 5.6 shows the mean (white circle) and credible intervals (blue vertical lines) of the spatial random effects estimated by the MCAR (left) and MDAGAR (right) models, respectively. From these results, it seems that the estimated random effects from both models are very similar. Although the majority of spatial random effects estimated by both models are not significantly different from zero, which is natural due to the distribution of the spatial random effects, most of the posterior mean estimates are different from zero.

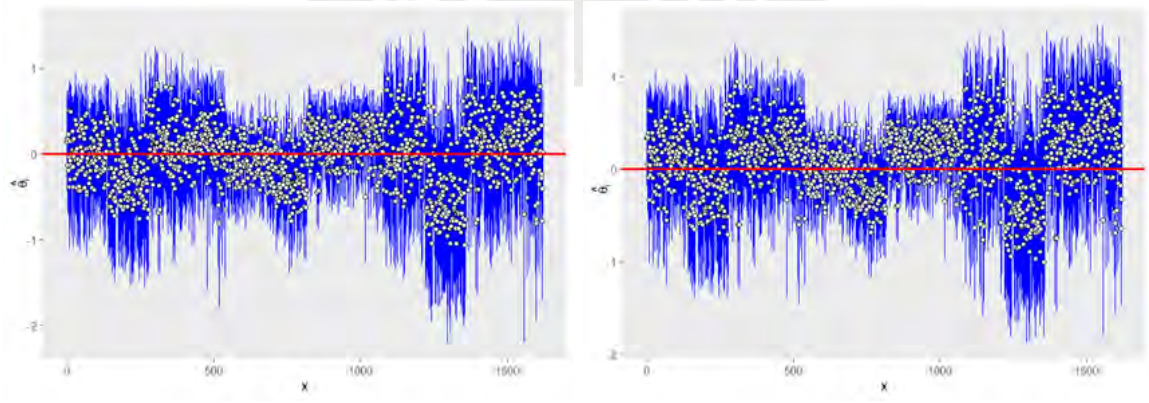


Figure 5.6: Posterior means (white circles) and 95 % credible intervals (blue) of the spatial random effects estimated by MCAR (left) and MDAGAR (right) models.

Figure 5.7 shows maps of the posterior means of the spatial random effects estimated by both models. In both models, the means of spatial random effects in the central east and south present mostly positive effects while the negative effects are concentrated in the north west and central west. These maps also allow us to observe the main difference between the spatial distributions estimated across the municipalities through the MCAR and MDAGAR models. For instance, in the oral cancer, the spatial random effects in the north are higher through the MDAGAR models than through the MCAR models.

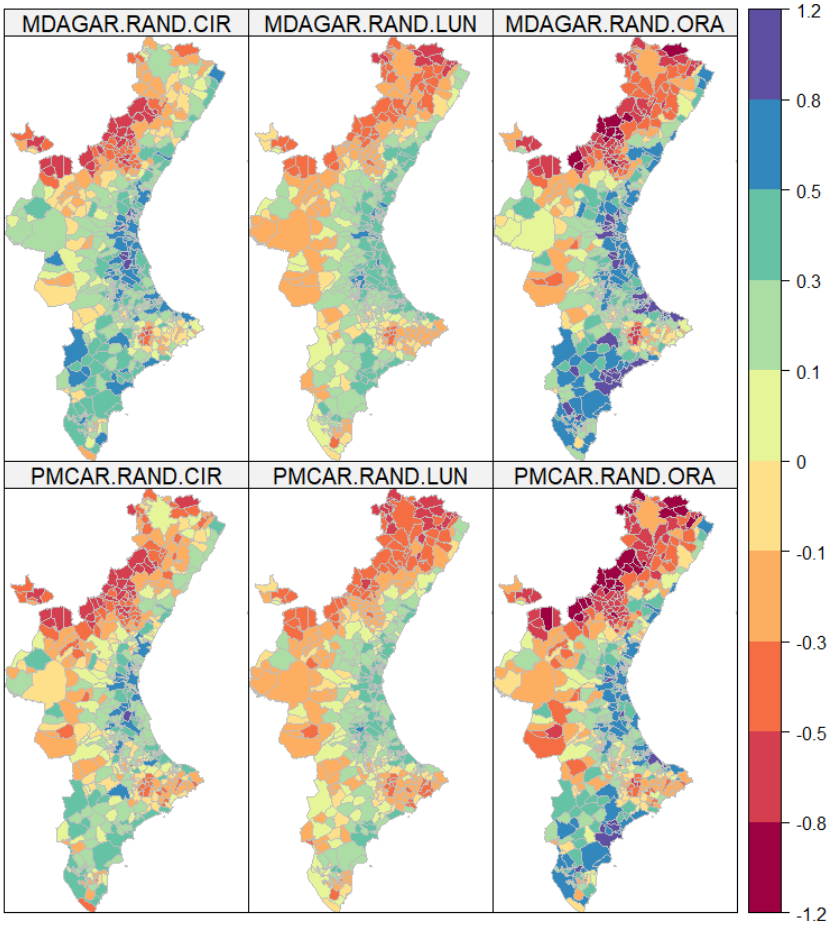


Figure 5.7: Maps of the spatial random effects estimated by MCAR (bottom) and MDAGAR (top) models for each disease in Comunidad Valenciana.

The relative risks estimated by both models follow a similar pattern across diseases with higher relative risks grouped in the central east coast and south as can be observed in Figure 5.8. Figure 5.9 shows the relationship between the actual relative risks and the estimations done by the MCAR model. As can be seen, the relative risk of some municipalities takes values as large as 10, nevertheless, the estimations are never greater than 2. This might be due to the lack of covariates in the model.

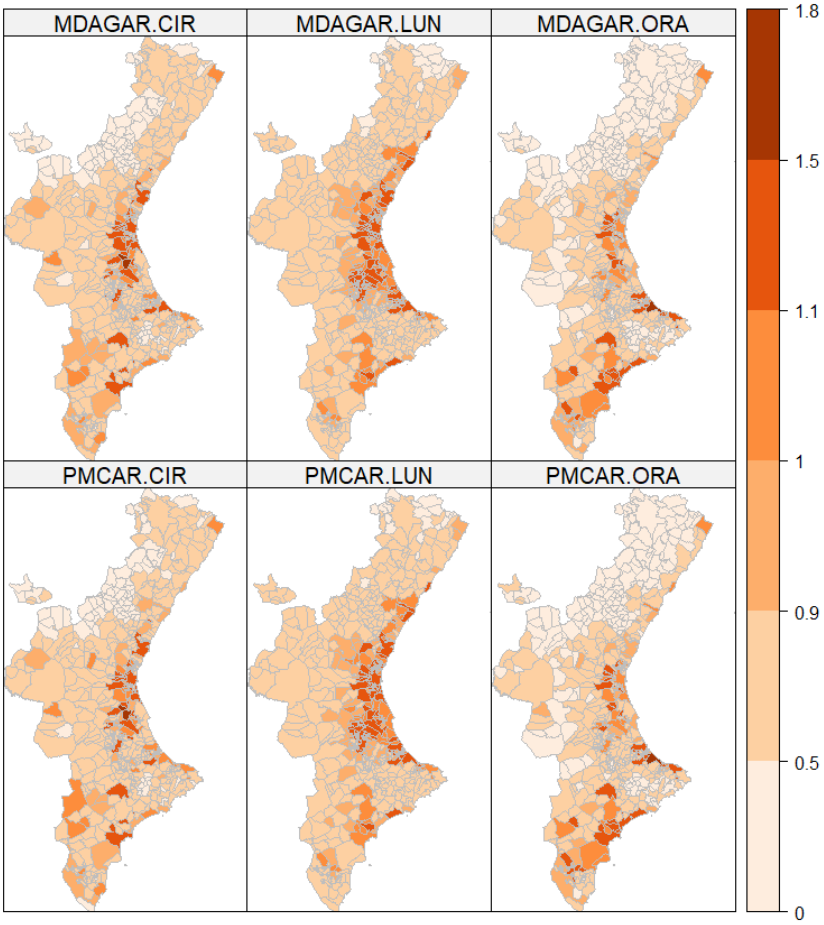


Figure 5.8: Maps of the posterior means of the relative risk estimated by MCAR (bottom) and MDAGAR (top) models for each disease in Comunidad Valenciana.

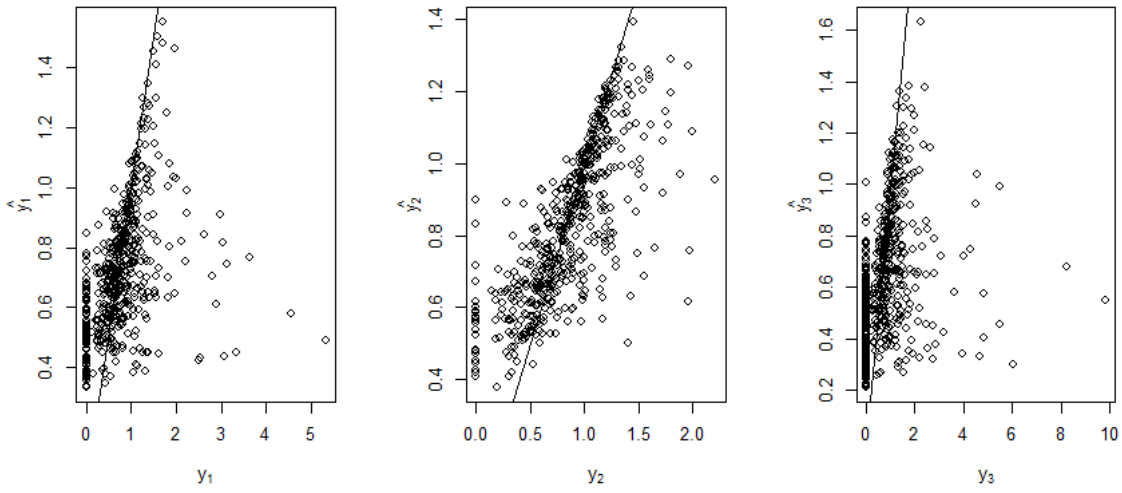


Figure 5.9: Plots of the real relative risk against the estimations done by MCAR model for each disease in Comunidad Valenciana.

5.2 Application 2: ARI and Anemia in Peru

In this application, we analyze the risk of acute respiratory infections (ARI) and anemia in children under five years of age in 187 provinces of Peru in the year 2021. The standardized incidence ratios (SIR) were calculated as the proportion of observed counts with respect to the expected counts, where the expected count in a province is equal to the population of children under five years of age in that province multiplied by the total number of cases in children under five years of age in Peru and divided by the total population of children under five years of age in Peru. Figure 5.10 shows the standardized incidence ratio in 187 provinces of Peru. We observe that both diseases have a higher incidence in the jungle region of Peru.

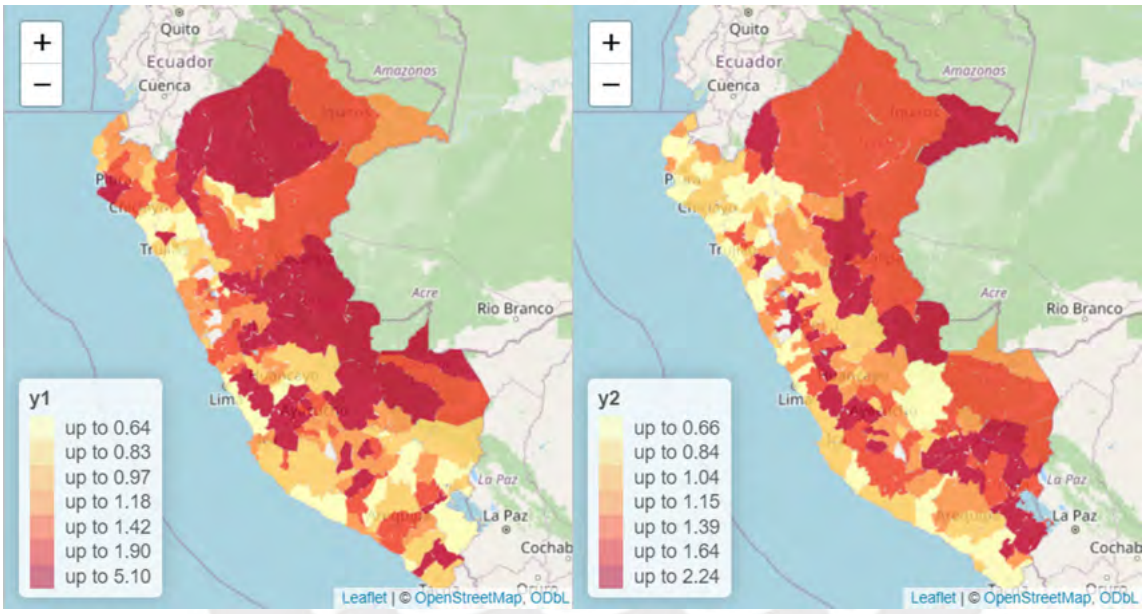


Figure 5.10: Maps of the standardized incidence ratio for ARI (left) and anemia (right) in the provinces of Peru.

An exploratory analysis of the spatial autocorrelation for the two diseases were performed using Moran's I, the following results were obtained, 0.306 for ARI and 0.319 for anemia.Since both value are higher than zero, we can conclude that there is evidence of spatial autocorrelation.

We assume that Y_{id} represents the observed number of children under five years old with ARI ($d = 1$) and anemia ($d = 2$) in province i, for $i = 1, 2, \ldots, n = 187$. The SIR of disease d in area i, R_{id}, is defined as the number of cases Y_{id} divided by the expected number of cases in children under five years old in that area E_{id}. We assume a multivariate Poisson distribution for Y_{id} such that:

$$Y_{id} \sim Poisson(\mu_{id} = E_{id}R_{id}); \quad i = 1, 2, \ldots, n = 187; \quad d = 1, 2,$$

where μ_{id} represents the average of cases of each disease d in the i-th province.

The mean is modeled through:

$$\log(\mu_{id}) = \log(E_{id}) + \beta_{0d} + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \theta_{id},$$

where β_{0d} is the intercept of the d-th disease, β_1 is the fixed effect that corresponds to the proportion of urban population (X_1), β_2 is the fixed effect that corresponds to the proportion of vaccinated people (X_2), β_3 is the fixed effect that corresponds to the elevation (X_3), β_4 is the fixed effect that corresponds to the precipitation (X_4), and θ_{id} represents the spatial random effect. We assume that these spatial random effects are defined from a MCAR and MDAGAR model. We then fit the MCAR and MDAGAR models through INLA.

Table 5.2 shows the posterior mean and 95% credible intervals for the parameters of each model, as well as the model selection criteria. The MDAGAR model shows better goodness of fit than the MCAR model according to the WAIC and LPML. In terms of estimation of the response variables, the MCAR model presents an slightly lower value of MSE. From these results, the MDAGAR model is selected as the best model.

Table 5.2: Mean and credible intervals (95%) for the estimated parameters of MCAR and MDAGAR models, as well as the model selection metrics.

Response	Parameter	MCAR	MDAGAR
y_1	β_{01}	0.386 (-0.022 ; 0.791)	0.187 (-0.2 ; 0.57)
	τ_1	1.059 (0.85 ; 1.3)	1.46 (0.713 ; 2.541)
y_2	β_{02}	0.357 (0.019 ; 0.69)	0.281 (-0.044 ; 0.602)
	τ_2	6.568 (4.321 ; 9.623)	9.794 (5.3 ; 16.143)
	β_1	-0.235 (-0.364 ; -0.105)	-0.251 (-0.374 ; -0.128)
	β_2	-0.613 (-0.975 ; -0.247)	-0.523 (-0.863 ; -0.178)
	β_3	0.789 (0.385 ; 1.193)	0.864 (0.427 ; 1.303)
	β_4	0.853 (0.402 ; 1.298)	0.771 (0.305 ; 1.234)
	ρ_{12}	0.157 (-0.067 ; 0.37)	0.142 (-0.098 ; 0.37)
	α	0.924 (0.825 ; 0.978)	0.697 (0.529 ; 0.844)
WAIC		3257.282	**3253.539**
LPML		-1874.288	**-1856.862**
MSE		**3.66E-02**	3.67E-02
Time (sec.)		6.89	15.1

Regarding the effect of the covariates, both models show similar results with the proportion of urban population and the proportion of vaccinated people having a negative effect on the SIR of both diseases and the elevation and precipitation of the province having a positive effects on the SIR of both diseases. According to the MDAGAR model, if there is an increment of one percentage point in the proportion of urban population, the SIR of both diseases decrease (1 - exp(-0.251))100%=22%. If there is an increment of one per-

centage point in the proportion of vaccinated people, the SIR of both diseases decrease (1-exp(-0.523))100%=41%. If there is an increment of one meter in elevation, the SIR of both diseases increase (exp(0.864)-1)100%=137%. If there is an increment of one unit in precipitation, the SIR of both diseases increase (exp(0.771)-1)100%=116%. The MCAR model estimated a very high mean α of 0.924 with credible interval [0.825, 0.978] suggesting a very high spatial autocorrelation, while the MDAGAR model gives a smaller estimate of the mean of $\alpha = 0.697$ with a wider credible interval of [0.529, 0.844] indicating a moderate to high spatial autocorrelation.

As can be seen in Figure 5.11, the marginal posterior distribution of precision hyperparameters from both models are different, with MDAGAR estimations indicating more precision and presenting more variance than the MCAR estimations. By contrast, the estimation of the hyperparameter of correlation between diseases are very similar for both models indicating a small correlation between ARI and anemia. Regarding the spatial autocorrelation parameter, the MCAR estimation presents a smaller variance than the MDAGAR estimation.

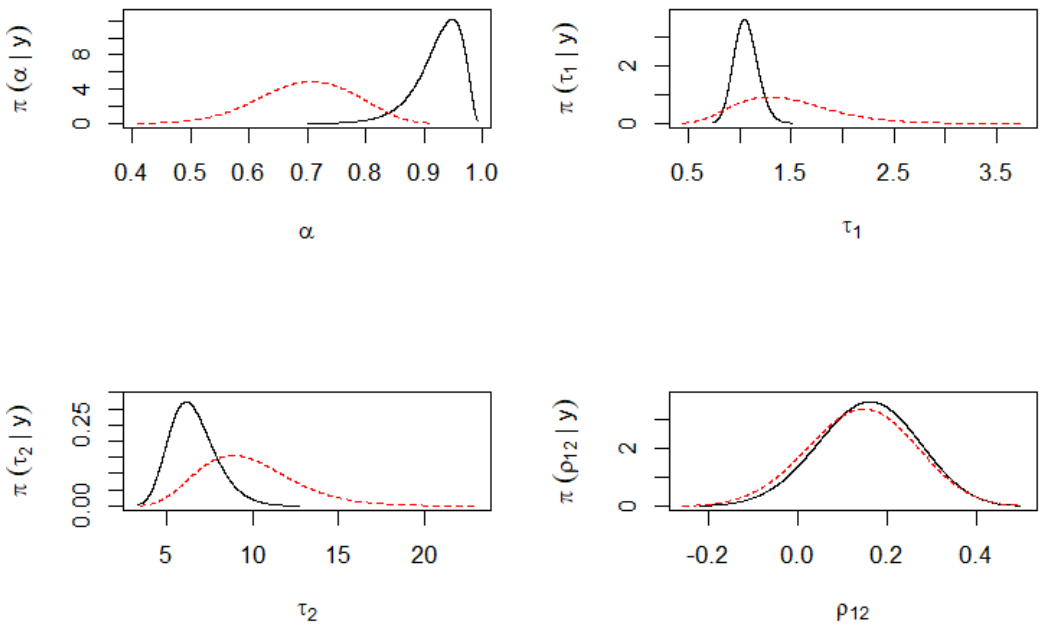


Figure 5.11: Marginal posterior density function of the hyperparameters estimated by MCAR (black) and MDAGAR (red) models.

Figure 5.12 shows the mean and 95% credible interval of the spatial random effects estimated by the MDAGAR model. Many of the spatial random effects estimated by the MDAGAR model for ARI are significantly different from zero.

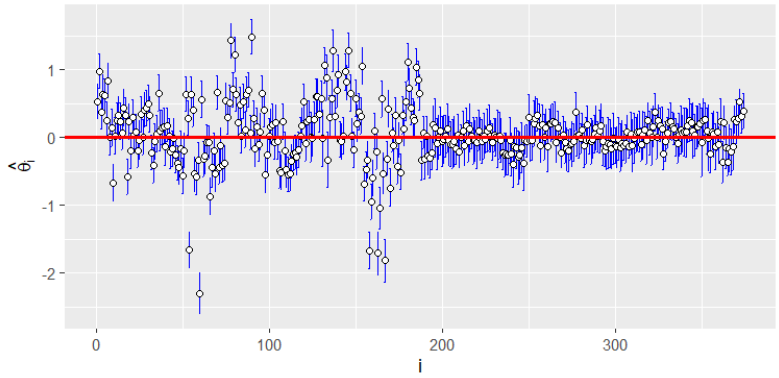


Figure 5.12: Means (white circles) and 95 % credible intervals (blue) of the spatial random effects estimated by MDAGAR model.

Figure 5.13 shows that each SIR estimated by both models have a similar spatial pattern. Furthermore, the mean posterior estimated risk is very similar to the true risk for each disease shown in Figure 5.10.

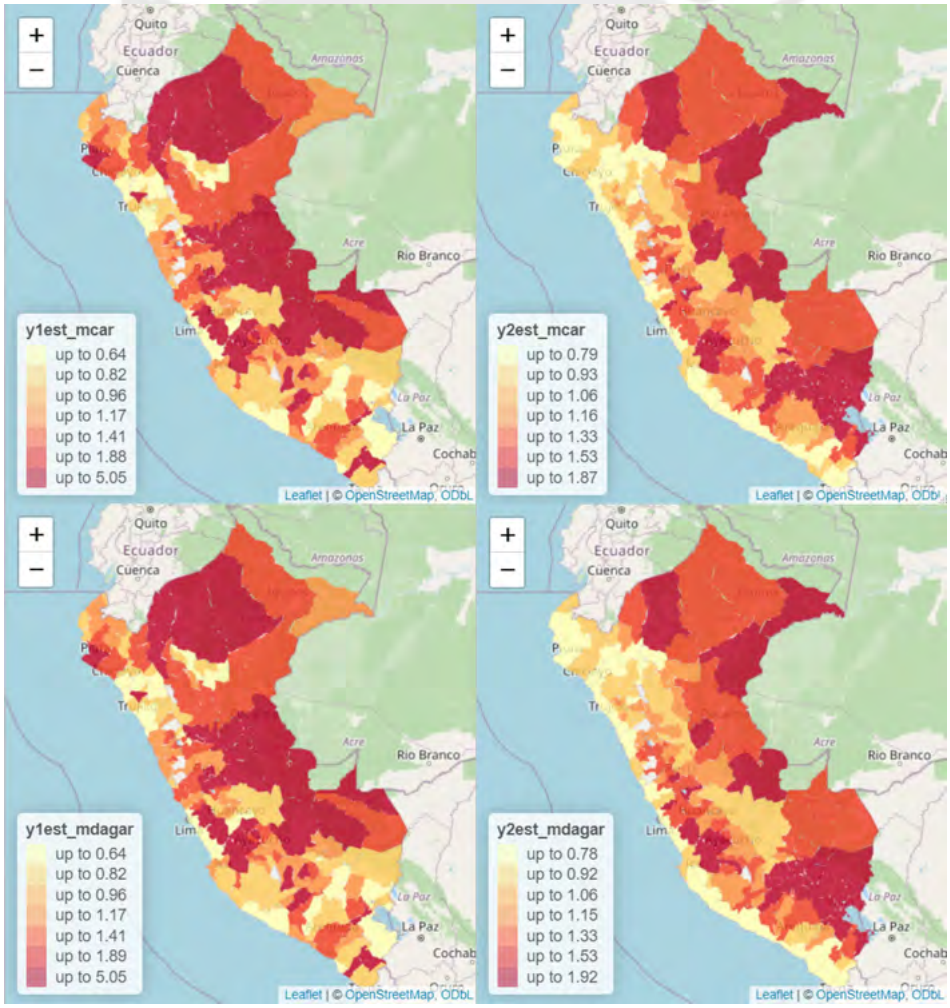


Figure 5.13: Maps of the posterior means of the standardized incidence ratio estimated by MCAR (top) and MDAGAR (bottom) models for each disease in Peru.

Figure 5.14 shows the relationship between the actual SIR and the estimations done by the MCAR and MDAGAR models. As can be seen, both models show similar estimates, with the estimates for IRA being much more precise than the estimations for anemia.

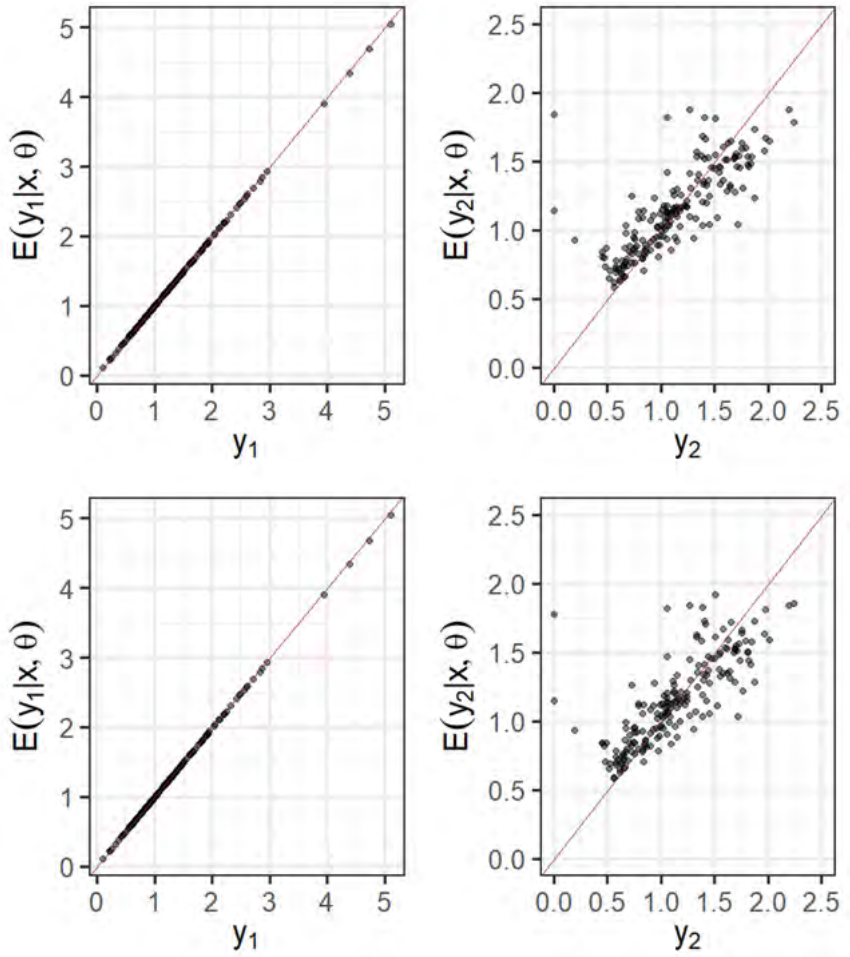


Figure 5.14: Plots of the real SIR against the estimations done by MCAR (top) and MDAGAR (bottom) models for each disease in Peru.

5.3 Application 3: Pneumonia, Anemia and EDA in Peru

In this application, we analyze the risk of pneumonia, anemia and acute diarrhoeal disease (EDA) in children under five years of age in 1814 districts of Peru in the year 2021. The standardized incidence ratios were computed for each disease. The SIR in a district was calculated as the proportion of observed counts in that district with respect to the expected counts for that district, where the expected count in a district is equal to the population of children under five years of age in that district multiplied by the total number of cases in children under five years of age in Peru and divided by the total population of children under five years of age in Peru. Figure 5.15 shows the standardized incidence ratio in 1814 districts

of Peru. It can be seen that the SIRs for pneumonia and EDA show similar spatial patterns with the highest SIRs clustered in the Peruvian Amazon and scattered in the Peruvian highlands. Regarding anemia, the highest SIRs are scattered in the Peruvian highlands and Amazon region. An exploratory analysis of the spatial autocorrelation for the three diseases were performed using Moran's I, the results obtained were 0.119 for pneumonia, 0.293 for anemia and 0.252 for EDA.

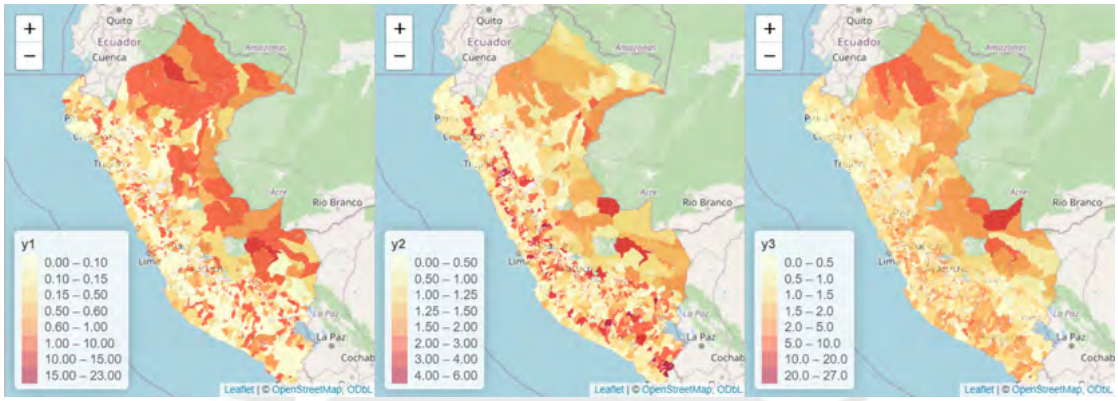


Figure 5.15: Maps of the standardized incidence ratio for pneumonia, anemia and EDA in the districts of Peru.

We assume that Y_{id} represents the observed number of children under five years old with pneumonia ($d = 1$), anemia ($d = 2$) and EDA ($d = 3$) in district i, for $i = 1, 2, \ldots, n = 1814$. The SIR of disease d in area i, (R_{id}), is defined as the number of cases (Y_{id}) divided by the expected number of cases in children under five years old in that area (E_{id}).

We assume a multivariate Poisson distribution for Y_{id} such that:

$$Y_{id} \sim Poisson(\mu_{id} = E_{id}R_{id}); \quad i = 1, 2, \ldots, n = 1814; \quad d = 1, 2, 3,$$

where μ_{id} represents the average of cases of each disease d in the i-th district. The mean is modeled through:

$$\log(\mu_{id}) = \log(E_{id}) + \beta_{0d} + \beta_1 x_{i1} + \theta_{id},$$

where β_{0d} is the intercept of the d-th disease, β_1 is the fixed effect that corresponds to the elevation in kilometers (X_1), and θ_{id} represents the spatial random effect. We assume that these spatial random effects are defined from MCAR and MDAGAR models, and fit these models through INLA.

Table 5.3 shows the posterior mean and 95% credible intervals for the parameters of each model, as well as the model selection criteria. The MDAGAR model shows better goodness of fit than the MCAR model according to the WAIC, LPML, and MSE. Regarding the effect

of the covariate, both models show similar results, with elevation having a positive effect on the SIR of the three diseases. Specifically, for each additional km of elevation, the SIR of each disease increases in 12.63%. The MCAR model estimated a very high posterior mean of α=0.932, with credible interval [0.909, 0.952], suggesting a very high spatial autocorrelation, while the MDAGAR model gives a smaller estimate of the posteriori mean of $\alpha = 0.252$ with a credible interval of [0.227, 0.278] indicating a small spatial autocorrelation.

Table 5.3: Mean and credible intervals (95%) for the estimated parameters of MCAR and MDAGAR models, as well as the model selection metrics.

Response	Parameter	MCAR	MDAGAR
y_1	β_{01}	-1.503 (-1.708 ; -1.303)	-1.469 (-1.6 ; -1.341)
	τ_1	0.186 (0.165 ; 0.208)	0.723 (0.642 ; 0.81)
y_2	β_{02}	-0.297 (-0.431 ; -0.164)	-0.311 (-0.393 ; -0.229)
	τ_2	0.476 (0.441 ; 0.514)	1.902 (1.746 ; 2.066)
y_3	β_{03}	-0.274 (-0.403 ; -0.147)	-0.244 (-0.325 ; -0.163)
	τ_3	0.55 (0.509 ; 0.594)	2.017 (1.86 ; 2.181)
	β_1	0.119 (0.085 ; 0.153)	0.103 (0.077 ; 0.13)
	ρ_{12}	0.053 (-0.015 ; 0.12)	0.051 (-0.015 ; 0.117)
	ρ_{13}	0.309 (0.247 ; 0.369)	0.328 (0.268 ; 0.387)
	ρ_{23}	0.034 (-0.017 ; 0.085)	0.069 (0.019 ; 0.119)
	α	0.932 (0.909 ; 0.952)	0.252 (0.227 ; 0.278)
WAIC		35175.714	**35050.542**
LPML		-65872.666	**-61741.065**
MSE		6.52E-02	**6.06E-02**
Time (sec.)		68.2	1383

As can be seen in Figure 5.16, from the posterior marginal distribution of the spatial autocorrelation parameter, both models present similar variance. The posterior marginal distribution of precision hyperparameters from both models are different, with MDAGAR estimations indicating more precision than MCAR estimations. By contrast, the estimation of the hyperparameter of correlation between diseases are very similar with both models indicating a small to moderate correlation between pneumonia and EDA and small correlation between pneumonia and anemia, and anemia and EDA.

Figure 5.17 shows the mean and 95 % credible interval of the spatial random effects estimated by the MDAGAR model. Most of the mean posterior spatial random effects estimates for pneumonia have more variance than for anemia and EDA.

Figure 5.18 shows that each SIR estimated from both models have a similar spatial pattern. Furthermore, the posterior mean estimated risk is very similar to the true risk for each disease shown in Figure 5.15.

Finally, Figure 5.19 shows the relationship between the actual SIR and the posterior mean estimations done by the MCAR and MDAGAR models. As can be seen, both models show

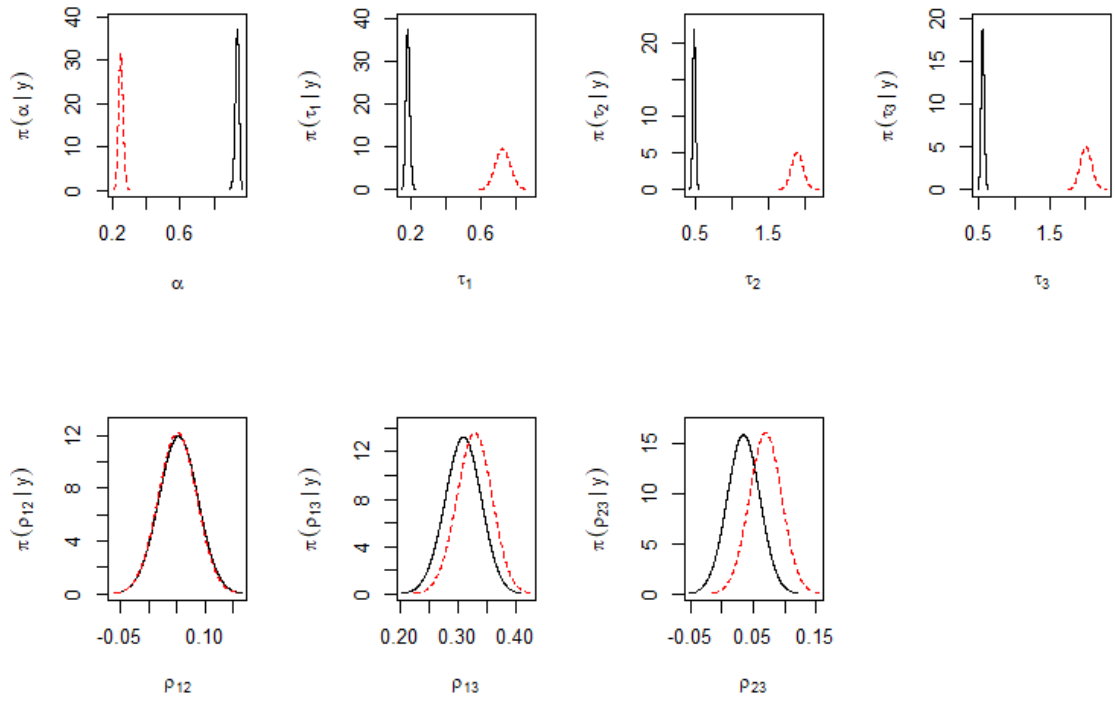


Figure 5.16: Probability density function of the hyperparameters estimated by MCAR (black) and MDAGAR (red) models.

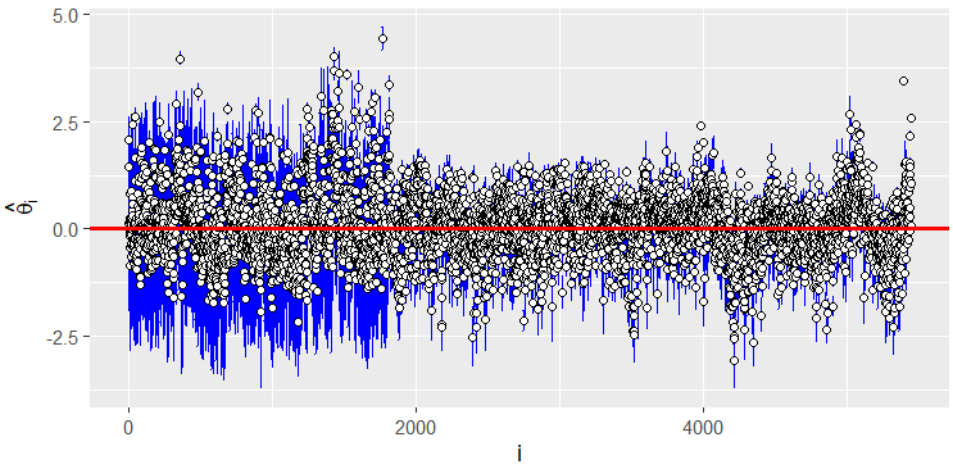


Figure 5.17: Means (white circles) and 95 % credible intervals (blue) of the spatial random effects estimated by MDAGAR model.

similar posterior mean estimates, with the estimates for pneumonia being less precise than the estimations for the other two diseases.

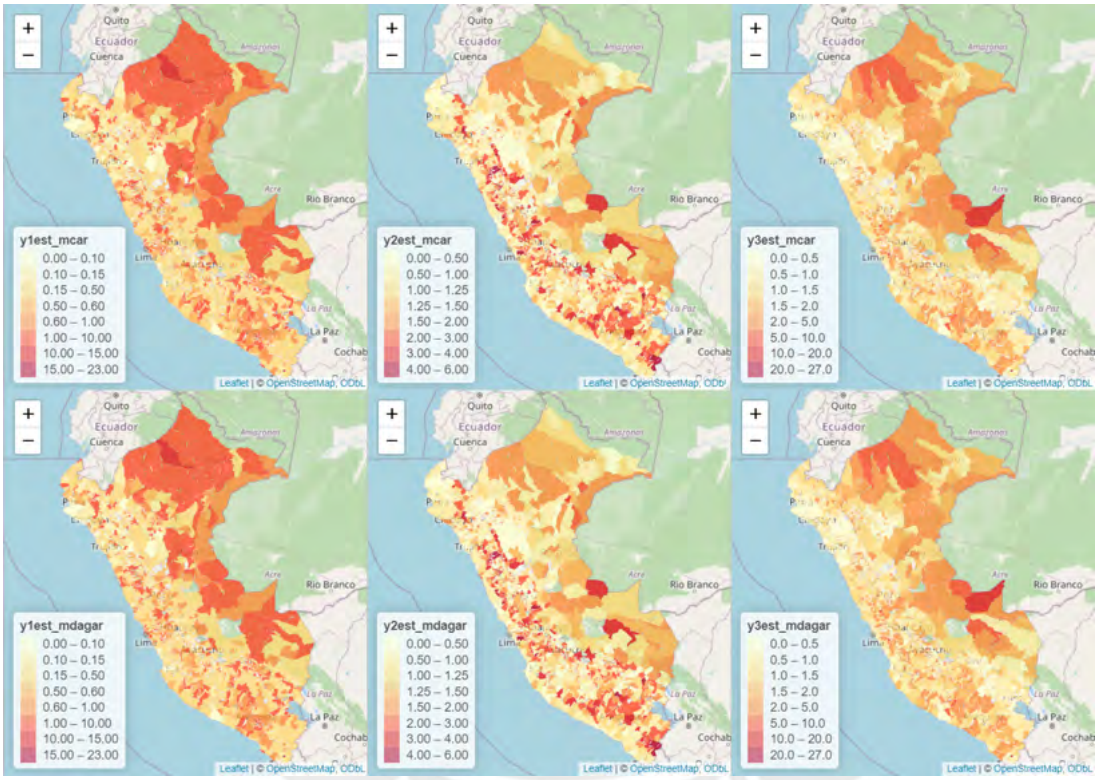


Figure 5.18: Maps of the posterior means of the standardized incidence ratio estimated by MCAR (top) and MDAGAR (bottom) models for each disease in Peru.

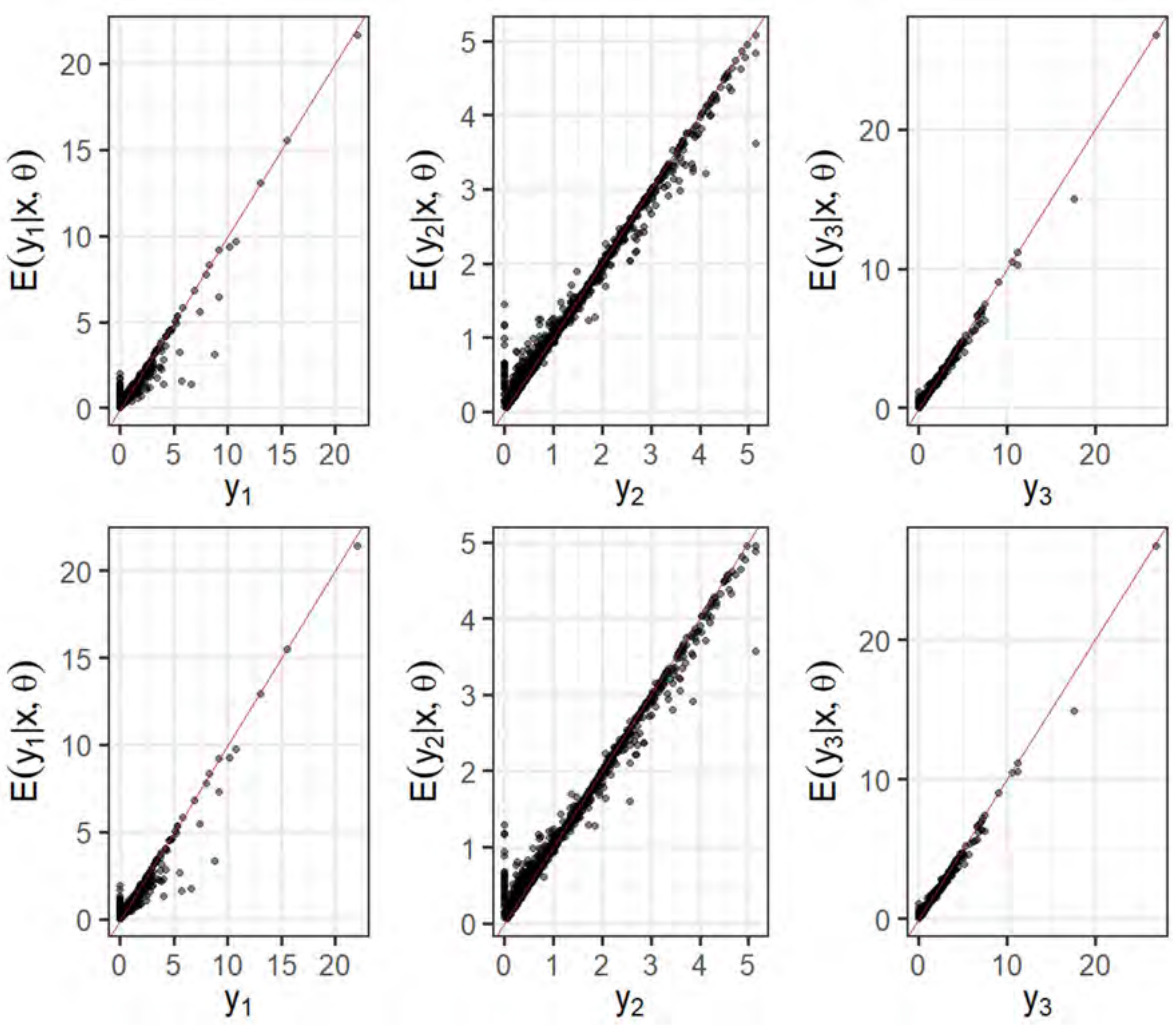


Figure 5.19: Plots of the real SIR against the posterior mean estimations through the MCAR model (top) and MDAGAR model (bottom) for each disease in Peru.

Chapter 6

Conclusions

This thesis introduced a spatial multivariate DAGAR (MDAGAR) model and its inference was implemented using the Integrated Nested Laplace Approximation (INLA) framework, aimed at more accurately capturing spatial autocorrelation in multivariate data settings. Through extensive simulation studies conducted under varying levels of spatial autocorrelation, from weak to strong, the proposed model demonstrated superior performance compared to the conventional Multivariate Conditional Autoregressive (MCAR) model, particularly in estimating spatial autocorrelation.

The robustness of the model was further validated using three real-world datasets, consistently outperforming MCAR across multiple goodness-of-fit metrics. Furthermore, the proposed model provided more accurate estimates of spatial autocorrelation than the MCAR model. These results not only reinforce the model's strength in detecting spatial dependencies but also highlight its practical utility in real-data applications.

In summary, the INLA-based spatial multivariate model presents a compelling alternative to traditional methods, offering enhanced estimation capabilities and improved model fit. These contributions pave the way for more nuanced and reliable spatial analyses, with promising implications for fields where spatial dependencies play a critical role such as epidemiology, ecology, and transportation where data-driven decisions can be taken for resource planning.

Future research could explore the extension of this modeling framework for the estimation of distinct spatial autocorrelation parameters for each variable within a multivariate context, thereby capturing variable-specific spatial structures with greater precision.

Bibliography

Adeyemi, R. A., Zewotir, T. and Ramroop, S. (2019). Joint spatial mapping of childhood anemia and malnutrition in sub-saharan africa: a cross-sectional study of small-scale geographical disparities, *African health sciences* **19**(3): 2692–2712.

Aswi, A., Cramb, S., Moraga, P. and Mengersen, K. (2019). Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review, *Epidemiology & Infection* **147**: e33.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, CRC Press.

Basseville, M., Benveniste, A., Chou, K. C., Golden, S. A., Nikoukhah, R. and Willsky, A. S. (1992). Modeling and estimation of multiresolution stochastic processes, *IEEE Transactions on Information Theory* **38**(2): 766–784.

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

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V. and Pebesma, E. J. (2008). *Applied spatial data analysis with R*, Vol. 747248717, Springer.

Cliff, A. D. and Ord, J. K. (1981). *Spatial processes: models and applications*, Pion Ltd, London, UK.

Cressie, N. (2015). *Statistics for spatial data*, John Wiley & Sons.

Datta, A., Banerjee, S., Hodges, J. S. and Gao, L. (2019). Spatial disease mapping using directed acyclic graph auto-regressive (DAGAR) models, *Bayesian Analysis* **14**(4): 1221–1244.
URL: *https://doi.org/10.1214/19-BA1177*

Geary, R. C. (1954). The contiguity ratio and statistical mapping, *The incorporated statistician* **5**(3): 115–146.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American statistical association* **85**(410): 398–409.

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis, *Biostatistics* **4**(1): 11–25.
URL: *https://doi.org/10.1093/biostatistics/4.1.11*

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, Vol. 580, Springer.

Huang, H., Zhou, H., Wang, J., Chang, F. and Ma, M. (2017). A multivariate spatial model of crash frequency by transportation modes for urban intersections, *Analytic methods in accident research* **14**: 10–21.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk, *Statistics in medicine* **19**(17-18): 2555–2567.

Lichstein, J. W., Simons, T. R., Shriner, S. A. and Franzreb, K. E. (2002). Spatial autocorrelation and autoregressive models in ecology, *Ecological monographs* **72**(3): 445–463.

Mardia, K. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing, *Journal of Multivariate Analysis* **24**(2): 265–284.
URL: *https://www.sciencedirect.com/science/article/pii/0047259X88900401*

Martinez-Beneito, M. A. (2013). A general modelling framework for multivariate disease mapping, *Biometrika* **100**(3): 539–553.
URL: *https://doi.org/10.1093/biomet/ast023*

Martínez-Beneito, M. A. and Botella-Rocamora, P. (2019). *Disease mapping: from foundations to multidimensional modeling*, CRC Press.

Moran, P. A. (1950). Notes on continuous stochastic phenomena, *Biometrika* **37**(1/2): 17–23.

Palmí-Perales, F., Gómez-Rubio, V. and Martinez-Beneito, M. A. (2021). Bayesian multivariate spatial models for lattice data with INLA, *Journal of Statistical Software* **98**(2).

Rue, H., Martino, S. and Chopin, N. (2009). Approximate bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(2): 319–392.

Schmidt, A. M. and Nobre, W. S. (2018). *Conditional Autoregressive (CAR) Model*, John Wiley Sons, Ltd, pp. 1–11.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08048*

Tesema, G. A., Tessema, Z. T., Heritier, S., Stirling, R. G. and Earnest, A. (2023). A systematic review of joint spatial and spatiotemporal models in health research, *International Journal of Environmental Research and Public Health* **20**(7): 5295.

VanderWeele, T. J. and Robins, J. M. (2007). Four types of effect modification: a classification based on directed acyclic graphs, *Epidemiology* **18**(5): 561–568.

Wall, W. (2004). A close look at the spatial structure implied by the CAR and SAR models, *Journal of Statistical Planning and Inference* **121**(311-324).

Whittle, P. (1954). On stationary processes in the plane, *Biometrika* pp. 434–449.