Research, Society and Development, v. 10, n. 1, e31910111890, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i1.11890

Inferência Bayesiana aplicada em modelo de regressão linear e modelo espacial: Uma abordagem sobre a estrutura de covariância entre os dados geoestatísticos

Bayesian inference applied to linear regression model and spatial model: An approach on the covariance structure between geostatistical data

Inferencia Bayesiana aplicada al modelo de regresión lineal y al modelo espacial: Un enfoque sobre la estructura de covarianza entre datos geoestadísticos

Recebido: 11/01/2021 | Revisado: 12/01/2021 | Aceito: 14/01/2021 | Publicado: 14/01/2021

Rondinelli Gomes Bragança
ORCID: https://orcid.org/0000-0002-4195-9547
Universidade Federal Fluminense, Brasil
E-mail: rondinelli.123@gmail.com

Resumo
Modelos estatísticos servem para descrever o comportamento probabilístico de fenômenos de interesse permitindo analisá-los, prevê-los e tomar decisões pertinentes. Modelos de regressão linear são muito utilizados em diversas áreas. Esses modelos possuem suposições fortes como independência entre os erros que em geral não se ajustam a dados espaciais, já que estes dados permitem que haja dependência na estrutura de covariância dos erros. Portanto, modelos de regressão linear podem ser comparados com modelos espaciais. Dados espaciais podem ser divididos em 3 tipos: padrão de pontos, dados de área e dados geoestatísticos. Esse trabalho visa avaliar modelos de regressão linear inicialmente e posteriormente compará-los aos modelos espaciais para dados geoestatísticos através da função de covariância exponencial. Parâmetros desconhecidos são encontrados nesses modelos e a inferência adotada nesse trabalho é a Bayesiana por permitir que a crença inicial do especialista seja incorporada a modelagem, aumentando a quantidade de informação avaliada e melhorando portanto as estimativas. Ao ajustar os modelos sob conjuntos de dados simulados é possível verificar a capacidade dos ajustes recuperarem os verdadeiros valores dos parâmetros e selecionar o verdadeiro modelo. O presente artigo é resultado do interesse em analisar o ajuste do modelo de regressão linear com conjunto de dados artificiais com dependência espacial e comparar esse ao ajuste do modelo espacial, mais especificamente, a partir de dados geoestatísticos.
Palavras-chave: Estatística espacial; Geoestatística; Inferência Bayesiana; Modelo de regressão linear; Métodos de Monte Carlo via cadeias de Markov; DIC; Erro médio quadrático.

Abstract
Statistical models serve to describe the probabilistic behavior of phenomena of interest, allowing them to be analyzed, predicted and made relevant decisions. Linear regression models are widely used in several areas. These models have strong assumptions such as independence between errors that in general do not fit spatial data, since these data allow dependence on the error covariance structure. Therefore, linear regression models can be compared with spatial models. Spatial data can be divided into 3 types: dot pattern, area data and geostatistical data. This work aims to evaluate linear regression models initially and later to compare them to spatial models for geostatistical data through the exponential covariance function. Unknown parameters are found in these models and the inference adopted in this work is Bayesian for allowing the expert's initial belief to be incorporated into the modeling, increasing the amount of information evaluated and therefore improving the estimates. When adjusting the models under simulated data sets, it is possible to verify the ability of the adjustments to recover the true values of the parameters and select the true model. This article is the result of an interest in analyzing the adjustment of the linear regression model with a set of artificial data with spatial dependence and comparing this to the adjustment of the spatial model, more specifically, based on geostatistical data.
Keywords: Spatial statistics; Geostatistics; Bayesian inference; Linear regression model; Markov chains Monte Carlo; DIC; Mean square error.

Resumen
Los modelos estadísticos sirven para describir el comportamiento probabilístico de los fenómenos de interés, permitiendo analizarlos, predecirlos y tomar decisiones relevantes. Los modelos de regresión lineal se utilizan ampliamente en varias áreas. Estos modelos tienen fuertes supuestos como la independencia entre errores que en general no se ajustan a los datos espaciales, ya que estos datos permiten la dependencia de la estructura de covarianza del error. Por lo tanto, los modelos de regresión lineal se pueden comparar con modelos espaciales. Los datos espaciales se pueden

dividir en 3 tipos: patrón de puntos, datos de área y datos geoestadísticos. Este trabajo tiene como objetivo evaluar inicialmente los modelos de regresión lineal y posteriormente compararlos con modelos espaciales para datos geoestadísticos mediante la función de covarianza exponencial. En estos modelos se encuentran parámetros desconocidos y la inferencia adoptada en este trabajo es bayesiana para permitir que la creencia inicial del experto se incorpore al modelado, aumentando la cantidad de información evaluada y por tanto mejorando las estimaciones. Al ajustar los modelos bajo conjuntos de datos simulados, es posible verificar la capacidad de los ajustes para recuperar los valores reales de los parámetros y seleccionar el modelo verdadero. Este artículo es el resultado de un interés en analizar el ajuste del modelo de regresión lineal con un conjunto de datos artificiales con dependencia espacial y compararlo con el ajuste del modelo espacial, más específicamente, basado en datos geoestadísticos.
Palabras clave: Estadística espacial; Geoestadística; Inferencia Bayesiana; Modelo de regresión lineal; Métodos de Monte Carlo a través de cadenas de Markov; DIC; Error cuadrático medio.

1. Introdução

Modelos estatísticos consistem em atribuir afirmações probabilísticas sobre fenômenos de interesse e costumam ser definidos através de 2 características: o espaço amostral e a família de distribuições de probabilidades associadas. Após dados serem observados é possível avaliar se essa família ajustou-se satisfatoriamente aos fenômenos de interesse e, em caso afirmativo, é possível tomar decisões e realizar previsões, por exemplo.

Geralmente essa família possui parâmetros desconhecidos e há a necessidade em estimá-los. Existem diversas formas de inferir sobre esses parâmetros. A inferência clássica é a uma das mais utilizadas, porém, a inferência Bayesiana permite atribuir uma distribuição de probabilidade ao conjunto de parâmetros desconhecidos, na qual o especialista pode quantificar uma crença inicial sobre esse conjunto, tornando essa forma de inferência muito atrativa. A estimação desse conjunto é realizada através da distribuição a posteriori e essa costuma possuir forma analítica desconhecida. É possível recorrer aos métodos de simulação como os métodos de MCMC.

Um dos modelos estatísticos mais utilizado é o de regressão linear. Esse modelo descreve a relação entre duas ou mais variáveis e costuma ser fácil de ajustar e interpretar e, portanto, é muito utilizado em diversas áreas. Essa modelagem pode ser aplicada para relacionar a taxa de criminalidade com a taxa de desemprego ou a ocorrência de chuva com a temperatura e umidade ou a quantidade de pessoas infectadas com dengue com o nível econômico de uma região, por exemplo. Porém, esta modelagem requer algumas suposições importantes para implementação do modelo, que, na prática, é difícil de ser alcançada. Uma delas é a independência entre os erros. A estatística espacial permite atribuir uma estrutura de dependência na matriz de covariâncias dos erros e, assim, é possível modelar um experimento com esta natureza.

A estatística espacial tem por objetivo estudar quantitativamente dados de natureza espacial da seguinte forma: identificando, analisando e modelando a ocorrência desses fenômenos posicionados no espaço. Ao decorrer dos anos constatou-se que era necessário um avanço tecnológico para auxiliar no estudo desta área.

O avanço da informática e de tecnologias têm proporcionado um aumento na quantidade de dados armazenados e contribuído para o aperfeiçoamento da Estatística Espacial. Na década de 1960 observou-se a necessidade de um programa para armazenar e analisar a grande quantidade de dados espaciais, já que naquela época ainda se iniciava a era computacional. Diante dessa necessidade e logo após entre os anos de 70 e 80, foi imposto o SIG, do inglês *Geographic Information System (GIS)*, com objetivo de criar um inventário de recursos naturais, coletando, armazenando, manipulando, visualizando e analisando dados espacialmente referenciados a um sistema de coordenadas conhecido. Câmara e Ortiz (1998) definiu que este sistema deve ser capaz de integrar dados de informações espaciais como censo e cadastros, combinar os vários dados para gerar mapeamentos e plotar este conteúdo da base de dados geocodificados.

Com o intuito de estimar o nível de nitrato, um composto químico que ocorre naturalmente em águas subterrâneas, Woodard e O'Connell (2010) realizou um estudo nas águas dos Estados do meio Atlântico. Para esse estudo foi utilizada a análise espacial onde o interesse era estimar se a quantidade do nível de nitrato era superior a um determinado limiar. Este trabalho recorreu a inferência bayesiana e para a modelagem espacial utilizou-se a função de covariância esférica.

Outro exemplo de aplicação da estatística espacial pode ser dado na área da Epidemiologia Espacial, que é uma junção da estatística espacial à epidemiologia. Borgoni e Billari (2003) destacam que os argumentos da análise espacial podem fornecer informações úteis sobre as necessidades de saúde reprodutiva não atendidas. Esse trabalho avalia o uso de preservativos em mulheres em sua primeira relação sexual e conclui que o sul da Itália apresenta maiores índices de relações sexuais desprotegidas pela primeira vez. Um importante resultado para a sociedade, sustentando a tomada de decisão e - para quando se aplicar - a fomentação de políticas públicas para prevenir tal estatística.

Krige (1951) concluiu em suas análises que não era possível estimar adequadamente a quantidade de ouro que se encontrava em blocos de minério caso não fosse considerados a localização e o volume.

Neste trabalho, utilizaremos dados geoestatísticos com base em conjunto de dados gerados com intuito de fazer um estudo quantitativo e analisar o funcionamento de cada tipo de modelo ajustado, (Pereira, et al. (2018)). O termo Geoestatística refere-se a um dos tipos de dados da estatística espacial caracterizados por Cressie (1993). A partir dos anos 60, a Geoestatística tem sido aplicada em diversos estudos na área de Geociências como em Pesquisa e Avaliação Mineral, em Hidrogeologia, Cartografia, Geologia Ambiental e Geotecnia.

2. Metodologia

Este capítulo apresenta uma breve introdução dos métodos utilizados neste trabalho que consiste inicialmente em avaliar modelos de regressão linear. Usualmente há parâmetros desconhecidos nesses modelos e nesse trabalho recorre-se a inferência Bayesiana para estimá-los. Sendo assim, a Seção 2.1 apresenta uma revisão sobre essa forma de inferência e a Seção 2.1.2 uma revisão de MCMC. A Seção 2.2 contem uma revisão de modelos de regressão linear. Logo após, na Seção 2.3 há uma revisão de modelos lineares com estrutura espacial nos erros e, em particular, apresenta-se uma discussão sobre geoestatística na Subseção 2.3.1. Além disso, neste trabalho serão abordados diferentes modelos e, para compará-los, na Seção 2.4 há uma breve introdução sobre dois métodos de comparação de modelos, o DIC e o EQM.

2.1 Inferência Bayesiana

Seja Y uma variável aleatória definida em um espaço amostral Ω. Suponha que haja interesse em uma característica populacional desconhecida e relacionada à essa variável. Considere que seja possível determinar a distribuição dessa variável condicionada a um vetor paramétrico θ, denotando-a por $p(Y = y|\theta)$. Suponha ainda que o vetor paramétrico seja desconhecido e que, com base nessa distribuição, seja possível analisar a característica populacional desconhecida. Sendo assim, faz-se necessário inferir sobre esse vetor paramétrico, ou seja, fazer afirmações sobre θ. Para isso, pode-se utilizar um conjunto de dados.

Sob a abordagem Bayesiana, é permitido incorporar uma crença inicial sobre θ, anterior à amostragem dos dados. Denote por $p(\theta)$ a distribuição a priori de θ que representa probabilisticamente essa crença inicial. A inferência sob θ é realizada com base na sua distribuição a posteriori, denotada por $p(\theta|Y = y)$, que é obtida através do Teorema de Bayes e dada pela seguinte equação:

$$p(\theta|Y = y) = \frac{p(Y = y|\theta)\, p(\theta)}{p(Y = y)} \tag{2.1}$$

A distribuição $p(Y = y)$ é chamada de distribuição marginal de Y e pode ser obtida combinando a distribuição $p(Y = y|\theta)$ com a distribuição $p(\theta)$.

Seja $c^{-1} = p(Y = y)$, então a distribuição a posteriori dada na Equação 2.1 pode ser reescrita como

$$p(\theta|Y = y) = c\ p(Y = y|\theta)\ p(\theta). \tag{2.3}$$

Note que a constante c não depende de θ. Por isso, sob a inferência Bayesiana, é comum utilizar a ideia de proporcionalidade e reescrever a Equação 2.1 da seguinte forma:

$$p(\theta|Y = y) \propto p(Y = y|\theta)\ p(\theta). \tag{2.4}$$

Desse modo, a expressão matemática costuma ser simplificada e torna-se mais fácil reconhecer o núcleo de uma distribuição conhecida, por exemplo. A constante c pode ser calculada recorrendo ao fato de que a integral (ou o somatório) de $p(\theta|Y = y)$ com respeito a θ tem que ser igual a 1.

Quando o vetor paramétrico θ for desconhecido, ao calcular a distribuição $p(Y = y|\theta)$ para um valor observado y da variável aleatória Y , obtém-se uma função que depende de θ. Essa função que é chamada de função de verossimilhança e passa a ser denotada por $l(y; \theta)$. Note que a função de verossimilhança não é uma função de distribuição. A integral de uma função de verossimilhança com respeito a θ pode ser diferente de 1, por exemplo. Essa expressão quando aplicada a diferentes valores de θ informa quais valores parecem ser mais verossímeis.

Em Migon, Gamerman e Louzada (2014) há maiores detalhes sobre inferência clássica e/ou Bayesiana, enquanto Casella e Berger (2002) contém informações sobre inferência clássica e Paulino et al. (2018) contém informações sobre análise Bayesiana.

Na Subseção 2.1.1 há uma discussão sobre como definir a distribuição a priori e a Subseção 2.1.2 apresenta um método iterativo que pode ser utilizado para avaliar a distribuição a posteriori quando essa for desconhecida.

2.1.1 Distribuição a priori

A distribuição a priori $p(\theta)$ deve representar toda a crença probabilística sobre o parâmetro de interesse θ. O especialista pode ter muito conhecimento prévio sobre o parâmetro desconhecido, tornando mais simples a tarefa de especificar uma distribuição a priori. Porém, ele também pode ter pouco conhecimento sobre θ e saber especificar apenas a média e a variância, por exemplo. Ou, ainda, pode não haver qualquer informação sobre θ antes do experimento.

Quando há informação sobre a distribuição a priori, pode-se recorrer a uma distribuição a priori conjugada e incorporar o conhecimento da média e da variância, se houver, através dos parâmetros dessa distribuição, chamados de hiperparâmetros. Caso não haja informação alguma, basta atribuir uma variância grande o suficiente para essa distribuição. Essa etapa será melhor explicada na Subseção 2.1.1.1.

Outro modo de atribuir uma distribuição a priori para um vetor paramétrico quando não há qualquer conhecimento prévio é o de recorrer a uma distribuição a priori não informativa e então a Subseção 2.1.1.2 apresenta uma revisão dessa distribuição.

2.1.1.1 Distribuição a priori Conjugada

A distribuição a priori do parâmetro desconhecido θ será definida através da distribuição da variável de interesse Y , $p(Y = y|\theta)$. Considere a seguinte definição (Ehlers (2003)):

Definição 2.1 Se $F = \{p(Y = y|\theta), \theta \in \Theta\}$ é uma classe de distribuições amostrais, então uma classe de distribuições P é conjugada a F se

$\forall\ p(Y = y|\theta) \in F$ e $p(\theta) \in P \Rightarrow p(\theta|Y = y) \in P$.

2.1.1.2 Distribuição a priori não informativa

As distribuições a priori podem ser classificadas de acordo com a informação probabilística dada, ou seja, não informativa ou informativa. O primeiro caso ocorre quando se assume ignorância probabilística total em relação ao parâmetro. Por exemplo, se um parâmetro assume valores no intervalo unitário e atribui-se uma distribuição a priori uniforme nesse intervalo, então todos os pontos possuem a mesma probabilidade de ocorrerem e portanto tem-se uma distribuição a priori não informativa. Porém, se o parâmetro assume valores na reta, por exemplo, não há uma distribuição uniforme que possa ser definida nesse intervalo. Nesses casos, há algumas alternativas. Primeiramente, pode-se tornar a distribuição a priori conjugada em uma distribuição muito ou pouco informativa escolhendo hiperparâmetros adequados, ou seja, quanto menor a variância dessa distribuição, maior será a informação e quanto maior for essa variância, mais vaga será a distribuição e portanto menos informativa. Outra forma, é recorrendo a distribuições a priori de referência e maiores detalhes podem ser encontrados em O'Hagan e Kendall (1994) e Bernardo e Smith (1994).

Em geral, as distribuições a posteriori dos parâmetros desconhecidos possuem forma analítica desconhecidas. Portanto, recorre-se aos métodos os MCMC para obter amostras destas distribuições. Neste trabalho em particular, serão aplicados o Amostrador de Gibbs e o algoritmo de Metropolis-Hastings. As Subseções seguintes apresentam uma revisão sobre os métodos MCMC e dos algoritmos citados.

2.1.2 Monte Carlo via Cadeias de Markov

Os métodos de Monte Carlo via Cadeias de Markov (MCMC) consistem em uma classe de algoritmos para amostrar de uma distribuição de probabilidade de interesse usando cadeias de Markov, sendo uma alternativa aos métodos não iterativos nos problemas em que as soluções analíticas tornam-se inviáveis ou complexas. Na inferência Bayesiana, os métodos de MCMC são muito utilizados para obter uma amostra da distribuição a posteriori de θ, permitindo assim inferir sobre o vetor paramétrico desconhecido. Como base na amostra obtida pelo MCMC, pode-se calcular as estimativas amostrais da distribuição de interesse. Uma discussão mais profunda pode ser encontrada em Robert e Casella (2004), Gamerman (2004) e Gamerman e Lopes (2006).

2.1.2.1 Cadeias de Markov

Uma cadeia de Markov de primeira ordem, ou processo de Markov de primeira ordem, é um processo estocástico $\{\theta 0, \theta 1, \ldots\}$ com espaço de estados finito ou infinito enumerável, tal que a distribuição de θt dado $\theta 0, \ldots, \theta t-1$ só depende do $\theta t-1$, ou seja,

$$p(\theta t \in A|\theta 0, \ldots, \theta t-1) = p(\theta t \in A|\theta\ t-1), \quad (2.5)$$

para qualquer subconjunto A. Isto é, a probabilidade de uma cadeia assumir um certo valor futuro, dado o estado atual e o estado passado, depende apenas do seu estado atual. De acordo com Gamerman e Lopes (2006), para que sejam usados os métodos MCMC, a cadeia deve ser homogênea (as probabilidades de transição de um estado para outro são invariantes), irredutível (cada estado pode ser atingido a partir de qualquer outro em um número finito de iterações) e aperiódica (não haja estados absorventes). Os métodos de MCMC mais utilizados na inferência Bayesiana são o amostrador de Gibbs e o algoritmo de Metropolis-Hastings. As Subseções seguintes fazem uma revisão sobre cada um desses algoritmos aplicados ao contexto da inferência Bayesiana.

2.1.2.2 Amostrador de Gibbs

O Amostrador de Gibbs, proposto por Geman e Geman (1984) e introduzido por Gelfand e Smith (1990), é uma cadeia de Markov na qual não há um método de aceitação- rejeição, ou seja, a cadeia sempre irá para um novo valor.

Considere que o interesse esteja em amostrar um vetor ou matriz θ da distribuição a posteriori p(θ|Y = y), sendo y um conjunto de dados observados. Suponha que esse conjunto θ seja particionado em d componentes e que cada componente possa ser um escalar ou um vetor ou mesmo uma matriz. Por simplicidade, considere que sejam d escalares. O amostrador de Gibbs requer a obtenção das distribuições condicionais completas a posteriori, ou seja, das distribuições p(θi|θt−i, y), para i = 1, . . . , d e θt−i = (θ1, . . . , θi− 1, θi+1, . . . , θd) (vetor com elemento θi excluído). As transições de um estado para o outro são feitas através dessas distribuições. Após a convergência, os valores resultantes formam uma amostra de p(θ|Y = y).

Espera-se que a convergência das cadeias de Markov seja atingida após um número de iterações suficientemente grande e após o período de aquecimento. O período de aquecimento corresponde as iterações iniciais necessárias até o momento em que a cadeia começa a convergir e costuma ser facilmente identificado ao analisar graficamente o traço da cadeia do parâmetro amostrado.

Algo importante a ser observado é o fato de que os parâmetros amostrados podem ser altamente autocorrelacionados. Uma solução para esse problema é o uso de espaçamento de ordem k para que seja selecionada uma amostra a cada k iterações, corrigindo assim a autocorrelação da cadeia. O valor de k pode ser estimado através do gráfico da autocorrelação do parâmetro amostrado. Um outro recurso muito utilizado é rodar o algoritmo r vezes alterando os valores iniciais dos parâmetros em cada rodada e comparando se após a convergência as diferentes amostras de cada parâmetro pertencem a mesma região.

2.1.2.3 Algoritmo de Metropolis-Hastings

Os algoritmos de Metropolis-Hastings, introduzido por Metropolis (1953) e estendido por Hastings (1970) para o caso mais geral, tem por objetivo simular uma distribuição de probabilidade desconhecida. Este algoritmo nos garante a convergência para uma certa distribuição, chamada de distribuição de equilíbrio, que, na inferência Bayesiana, pode ser a distribuição a posteriori.

A escolha da distribuição proposta $q(\theta^p|\theta^{t-1})$ é fundamental para o funcionamento desse algoritmo. A distribuição proposta, se tratando como passeio aleatório, pode ser escolhida arbitrariamente, mas na prática deve-se tomar alguns cuidados para garantir a eficiência do algoritmo. Se atribuir um valor pequeno para a variância da proposta, a convergência da Cadeia de Markov será lenta, já que seus incrementos serão pequenos. No entanto, se a variância assumir valores muito altos, a cadeia tenderá a não se mover, já que a taxa de rejeição dos valores propostos será alta.

A taxa de aceitação, que contabiliza quantas mudanças de estados ocorreram, depende da distribuição proposta. Essa taxa pode ser obtida dividindo o contador cont pelo total de iterações, e deve manter-se em torno de 0, 44, que é um valor ótimo para propostas unidimensionais (Roberts, Gelman & Gilks, 1997; Roberts & Rosenthal, 2001). Quando a taxa de aceitação for muito alta ou muito baixa, será necessário mudar a variância da distribuição ou até mesmo optar por outra distribuição proposta.

O algoritmo de Metropolis-Hastings utiliza os mesmos critérios de convergência vistos na subseção 2.1.2.2, sendo necessário um período de aquecimento e, possivelmente, o uso de espaçamento de ordem k a fim de evitar amostras altamente autocorrelacionadas. Maiores detalhes sobre estes e outros algoritmos relacionados podem ser obtidos, por exemplo, em Robert e Casella (2004) e Gamerman (2004). Um dos modelos que serão aplicados a este trabalho é o de regressão linear. Portanto, na Seção 2.2 fará uma breve revisão sobre este modelo.

2.2 Modelo de Regressão Linear

A regressão linear simples tem por objetivo verificar a existência de uma relação entre duas variáveis como, por exemplo, quando deseja-se investigar se o lucro obtido com determinado produto está relacionado com a renda média familiar de uma dada região. A variável de interesse é chamada de variável resposta ou variável dependente e a outra variável é chamada de explicativa ou covariável ou independente. Quando há mais de uma variável explicativa diz-se ter um modelo de regressão linear múltiplo e, nesse caso, o interesse pode estar em explicar o lucro através da renda, do valor gasto em propaganda e da densidade populacional da região, por exemplo.

Representa-se a relação entre variáveis através de um modelo estatístico e quando verifica-se a existência dessa relação, análise dos dados e previsões podem ser realizadas, auxiliando, por exemplo, em tomadas de decisões.

Considere n unidades amostrais como, por exemplo, indivíduos, países, animais, entre outros. Sejam Yi a variável resposta da i-ésima unidade amostral e Xi = (Xi0, . . . , Xi(p−1))'um vetor contendo as p variáveis explicativas da i-ésima unidade, em que Xi0 considera-se um vetor sendo todas as componentes iguais a 1. Suponha que, dado Xi, as variáveis respostas $Y_i = X_i' \beta + ei$ (2.6) sejam iid e ei segue distribuição normal com média zero e variância σ2, para todo i = 1, . . . , n, onde o vetor β = (β0, . . . , βp−1)' representa os efeitos das variáveis explicativas na variável resposta e os termos e1, . . . , en são considerados erros aleatórios.

Portanto, as suposições do modelo de regressão linear, com base na Equação 2.6, são:

1. O vetor de variáveis independentes Xi é conhecido para todas as unidades amostrais.
2. Linearidade: Dada as variáveis explicativas Xi, o valor médio da variável resposta é uma função linear dos parâmetros β.
3. Independência: Dada as variáveis explicativas Xi, as variáveis respostas são Independentes.
4. Normalidade: Os erros ei (i = 1, . . . , n) possuem distribuição normal conforme consta na Equação 2.6.
5. Homocedasticidade: A variância do erro ei é constante para todas as unidades amostrais

Em geral, assume-se que $X_{i0} = 1$ para todas as unidades amostrais e, nesse caso, β_0 é chamado de intercepto ou coeficiente linear e representa o valor médio de todas as unidades quando não há qualquer variável explicativa influenciando no processo. Supondo a existência de intercepto, quando p = 2 é dito ter um modelo de regressão linear simples e então β_1 é chamado de coeficiente angular e representa o efeito da variável explicativa na variável resposta.

Sob a abordagem Bayesiana, o vetor paramétrico desconhecido $\theta = (\beta, \sigma^2)$ pode ser estimado através da sua espectiva distribuição a posteriori que é dada combinando a função de verossimilhança, definida pela Equação em 2.6 com a distribuição a priori atribuída ao vetor paramétrico desconhecido.

2.2.1 Modelo de regressão linear proposto

Considere o modelo de regressão linear definido pela Equação em 2.6 e suponha que uma amostra de tamanho n foi observada. As variáveis respostas observadas serão denotadas pelo vetor y = (y1, . . . , yn)' e as variáveis explicativas serão representadas pela matriz x, onde cada linha corresponde ao vetor xi ' = (xi0, . . . , xi(p− 1)) observado para a i-ésima unidade.

Suponha a priori que β e σ2 sejam independentes, tais que o vetor de coeficientes β segue uma distribuição normal p-variada com vetor de média zero e matriz de covariância Σβ conhecida, e o parâmetro de precisão, definido por τ = σ-2 , segue uma distribuição Gama com parâmetros aτ e bτ conhecidos. Como essa distribuição multivariada é desconhecida, será necessário recorrer aos métodos de MCMC descritos na Subseção 2.1.2 para obter uma amostra dessa distribuição.

Esse modelo de regressão linear será aplicado nas Seções 3.1 e 3.2 do Capítulo 3. Esta última Seção compara dois tipos de ajuste de dados: o modelo de regressão linear e o modelo espacial. Portanto, na Seção 2.3 apresenta uma breve introdução sobre estatística espacial e o modelo espacial proposto, além do tipo de dado ao qual será abordado neste trabalho.

2.3 Estatística Espacial

Estatística Espacial é o conjunto de métodos estatísticos utilizados para analisar fenômenos que variam em um espaço geográfico considerando alguma estrutura espacial que há neste fenômeno estudado. Aplicações desta natureza são encontradas em diversas áreas como Epidemiologia, Estudos de violência, Agronomia, Demografia, Geologia, entre outros.

Cressie (1993) considera três tipos de dados para caracterizar os problemas de análise espacial: processos pontuais (ou padrões de pontos), dados de área e dados geoestatísticos (ou superfícies contínuas).

Processos pontuais, ou padrões de pontos, consistem em avaliar a ocorrência de um dado fenômeno considerando que a incerteza esteja na localização espacial dessa ocorrência. Por exemplo, suponha que o interesse seja prever a localização dos casos de furtos ocorridos em uma determinada cidade ou ainda que o interesse esteja em prever a localização de uma certa planta.

Para compreender sobre dados de área, suponha que a região espacial populacional seja dividida em áreas delimitadas como bairros, municípios ou países. Os fenômenos de interesse ocorridos em cada uma dessas áreas consistem no agrupamento/somatório dos fenômenos ocorridos em diferentes pontos dessa área. Um exemplo desse tipo de dados ocorre quando analisa-se o tempo de vida da população brasileira dividindo essa população em Estados. A informação de cada Estado consiste no agrupamento desses tempos em diferentes municípios. Um outro exemplo desse tipo de dados ocorre quando analisa-se o número de pessoas que contraíram dengue nos estados brasileiros e, nesse caso, os estados brasileiros correspondem as áreas delimitadas e o número de doentes em cada estado consiste a soma do número de doentes em diferentes ruas desse mesmo estado.

Na geoestatística a localização espacial é considerada conhecida e os dados não são agrupados. Por exemplo, suponha que o interesse esteja no teor de zinco no solo de uma dada região.

2.3.1 Geoestatística

A geoestatística, também nomeada de superfícies contínuas, estuda fenômenos que variam numa determinada região espacial em que as localizações são conhecidas.

De acordo com Cressie (1993), pode-se definir a região G como sendo subconjunto fixo do R_r com volume r-dimensional positivo, onde r = 1, 2 ou 3, usualmente. Seja $\{Y(s) : s \in G\}$ uma realização do processo aleatório. Com isso, pode-se dizer que s varia continuamente ao longo da região G e poderia conter informações como latitude, longitude e altitude.

Assuma que a média e a variância do processo $\{Y(s) : s \in G\}$ existam para todo $s \in G$. Então, denote-os por $\mu(s) = E(Y(s))$ e $V(Y(s))$, respectivamente. Considere que h é o vetor de separação. Define-se que o processo espacial $Y(\cdot)$ é intrinsecamente estacionário se

$$E[Y(s+h)] = E[Y(s)] \text{ e} \quad (2.10)$$

$$V[Y(s+h) - Y(s)] = E[(Y(s) - Y(s+h))^2] = 2\gamma(h), \quad (2.11)$$

para todo s, $s+h \in G$, em que a quantidade $\gamma(\cdot)$ é uma função condicionalmente negativa definida.

O processo $Y(\cdot)$ é dito estacionário de segunda ordem ou fracamente estacionário se $\mu(s)$ é constante para todo $s \in G$, ou seja, $\mu(s) = \mu$ para todo $s \in G$ e

$$cov\{Y(s), Y(s*)\} = C(h) \; \forall s, s* \in G, \quad (2.12)$$

isto é, a covariância entre dois pontos quaisquer em G é função apenas da diferença entre as duas localizações.

A quantidade $2\gamma(h)$ em (2.10) é um dos parâmetros mais importantes na modelagem de geoestatística e é conhecida como variograma e $\gamma(h)$ como semivariograma. Já a função $C(\cdot)$ dita em (2.12) é chamada de covariograma.

Quando o variograma depender apenas da distância entre as localizações s e $s*$, o processo é denominado por isotrópico. O processo se diz que é homogêneo quando o processo é intrinsecamente estacionário e isotrópico (Smith (1996)). O processo é heterogêneo se uma dessas duas condições não se aplicar.

Assume-se, em geral, que a variável $Y(\cdot)$, em que assume valores $y(s)$ para $s \in G$, segue um processo Gaussiano (PG) com média $\mu(\cdot)$ e função de covariância $c(\cdot, \cdot)$ onde é denotado por

$$Y(\cdot) \sim PG(\mu(\cdot), c(\cdot, \cdot)) \quad (2.13)$$

se para quaisquer $s_1, \ldots, s_n$, e qualquer $n = 1, 2, \ldots$, a distribuição conjunta de $Y(s_1), \ldots, Y(s_n)$ é uma normal multivariada com parâmetros dados por $E[Y(s_i)] = \mu(Y(s_i))$ e $cov\{Y(s_i, s_j)_i), Y(s_j)\} = c(s_i, s_j)$ (O'Hagan e Kendall (1994)).

Na Geoestatística a partir da estimação do semivariograma e construção do gráfico de semivariograma identifica-se a variabilidade espacial. Este permite representar graficamente a semivariância em função da distância, denominado de variograma experimental.

Ao representar graficamente um dado processo homogêneo, encontram-se três medidas que podem ser vistas por ele: o patamar, o efeito pepita e o alcance.

Quando o vetor de distância é nulo, por definição, o semivariograma é nulo, isto é, $\gamma(h) = 0$. Entretanto, na prática, o semivariograma amostral costuma ter valor não nulo quando $h = 0$. Este valor é chamado de efeito pepita e pode ser interpretado como um erro de medida.

O alcance é a distância a partir da qual as amostras passam a ser independentes, ou seja, quando não houver mais correlação espacial para o fenômeno de interesse.

O patamar é valor da variância que corresponde ao alcance e é considerado que a partir deste ponto não há mais a dependência espacial, já que a variância entre as amostras não varia em relação a distância.

No campo da geoestatística é necessário ajustar um modelo estatístico aos dados experimentais, em vista da necessidade de calcular o valor do semivariograma para qualquer distância. Para isso, existem diferentes funções de covariâncias válidas utilizadas na literatura assim como os modelos Exponencial, Gaussiano e esférico.

2.3.2 Modelo espacial proposto

Suponha que Y seja um vetor aleatório contínuo em uma região espacial G, ou seja, Y é a coleção de variáveis Y (s), s ∈ G. As coordenadas s podem ser a latitude, a longitude e altitude. No contexto da geoestatística assume-se que

$$Y \sim N_n(\mu, \Sigma) \quad (2.15)$$

m que N_n representa a distribuição normal multivariada de dimensão n, μ é um vetor que representa a média do processo com dimensão n e Σ representa a estrutura de covariância com dimensão n × n. O vetor aleatório Y descreve uma realização parcial do processo dado em 2.13 e, por isso, segue distribuição normal multivariada.

O vetor de médias μ é composto pela coleção de médias μ(s), s ∈ G e em geral pode ser descrito como $\mu(s) = X(s)^{'}\beta$ onde $\beta = (\beta_0, \ldots, \beta_{p-1})^{'}$ é um vetor de efeitos das covariáveis, sendo β_0 considerado como intercepto, e $X(s)^0 = (X_0(s), \ldots, X_{p-1}(s))$ um vetor contendo as p covariáveis do processo que explicam Y (s), em que $X_0(s)$ é um vetor que todas as componentes deste vetor é igual a 1. Considere d = h, sendo d a matriz de distâncias entre as localizações s e s∗.

Nesse trabalho o processo de interesse Y (s) é definido da seguinte forma

$$Y(s) = X(s)^{'}\beta + \omega(s) \quad (2.16)$$

onde ω(s) é um efeito espacial que segue distribuição normal com média 0, variância V para todo s e função de covariância dada por V × ρ(d), sendo ||d|| a distância euclidiana entre as localizações s e s∗ e ρ(d) a função de correlação espacial.

Então, a função de correlação exponencial $\rho(d) = \exp\{-d\nu\}$, onde ν > 0. Note que a componente espacial a ser estimada, ν, não entra diretamente no processo Y (s), mas na estrutura de correlação.

O vetor paramétrico a ser estimado será θ = (β, ν, V). Sob o enfoque Bayesiano faz-se necessário atribuir uma crença inicial a esse vetor. Em geral, assume-se que os parâmetros contidos em θ sejam independentes. Usualmente considera-se que o vetor de coeficientes contendo os efeitos das covariáveis possuem uma distribuição normal independente com média zero e variância $\sigma^2{}_\beta$, para i = 0, . . . , p − 1, sendo que σ_β é um valor conhecido e quanto maior, mais vaga é a informação inicial sobre β. Costuma-se utilizar uma distribuição gama inversa para os parâmetros de variância que equivale a utilizar uma distribuição gama para o parâmetro de precisão que é o inverso do parâmetro de variância. Além disso, costuma-se atribuir uma distribuição gama para o parâmetro ν que está relacionado com o alcance do processo.

Considere o modelo espacial definido pela Equação em 2.16 e suponha que uma amostra eorreferenciada de tamanho n foi observada nas localizações s em que s ∈ G. As variáveis respostas observadas serão denotadas pelo vetor $y(s) = (y_1(s), \ldots, y_n(s))^{'}$ e as variáveis explicativas serão representadas pela matriz x(s), onde cada linha corresponde ao vetor $x_i(s)^0 = (x_{i0}(s), \ldots, x_{i(p-1)}(s))$ observado para a i-ésima unidade.

Suponha a priori que β e σ^2 sejam independentes, tais que o vetor de coeficientes β segue uma distribuição normal p-variada com vetor de média zero e matriz de covariância Σ_β conhecida, e o parâmetro de precisão, definido por $\tau = \sigma^{-}{}_2$, segue uma distribuição Gama com parâmetros a_τ e b_τ conhecidos e Σ = V × ρ(d).

Como a distribuição a posteriori de (β, τ, ν) é desconhecida, será necessário recorrer aos métodos de MCMC descritos na Subseção 2.1.2 para obter uma amostra dessa distribuição.

Como a função de correlação exponencial nunca se anula, recorre-se a ideia de alcance efetivo que consiste em considerar que duas localizações não possuem dependência espacial quando sua função de correlação exponencial assumir um valor igual ou inferior a 0,05. E então o alcance h_a é dado por

$$\exp\{-h\nu\} \leq 0,05 \Leftrightarrow h \geq 3/\nu \Leftrightarrow h_a = 3/\nu \quad (2.19)$$

2.4 Comparação de modelos

Modelos estatísticos tentam descrever o comportamento probabilístico de um dado fenômeno de interesse e por isso diversos modelos podem ser utilizados em um mesmo conjunto de dados e posteriormente pode-se analisar qual modelo obteve melhor ajuste para descrever os dados observados e/ou prever futuros dados. Essa análise é realizada através da comparação de modelos e há vários critérios na literatura para isso. Esse trabalho recorrerá aos seguintes critérios: o critério de informação do desvio (DIC) e o erro quadrático médio (EQM). Ambos analisam a performance do modelo descrever os dados observados em termos de ajuste.

2.4.1 Critério de informação do desvio (DIC)

O critério de informação do desvio (DIC), proposto por Spiegelhalter et al. (2002), é uma generalização de outros métodos como o AIC e BIC e é utilizado para seleção de modelos Bayesianos em que a distribuição a posteriori é obtida pela simulação via MCMC.

Define-se o desvio da seguinte forma

$$D(\theta) = -2\log\{l(y;\theta)\} + C \quad (2.20)$$

sendo $l(y;\theta)$ a função de verossimilhança para os dados observados y e C uma constante.

O critério DIC é definido então como

$$DIC = 2D(\theta) - D(\hat{\theta}) \quad (2.21)$$

sendo $D(\theta)$ a média a posteriori dos desvios definidos na Equação (2.20) e calculado da seguinte forma $D(\theta) = Som_{it=1} D(\theta^{it})/m$, com m representando o total de iterações do m MCMC e θ^{it} o valor amostrado na it-ésima iteração. O termo $\hat{\theta}$ representa a média a posteriori dos parâmetros do modelo e C é uma constante que se cancela, portanto não é levado em consideração na comparação de modelos.

Esse critério pode ser reescrito como

$$DIC = p_D + D(\theta) \quad (2.22)$$

sendo $p_D = D(\theta) - D(\hat{\theta})$ considerado como o número efetivo de parâmetros no modelo.

Segundo esse critério, o melhor modelo terá o menor valor de DIC. Mais detalhes podem ser encontrados em Macera (2011).

2.4.2 Erro Quadrático Médio (EQM)

O EQM é definido como sendo a média do quadrado da diferença entre o valor do estimador e do parâmetro (Morettin e Bussab (2010)). Assim, para um estimador $\hat{\theta}$ do parâmetro θ, temos:

$$EQM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = V\,ar(\hat{\theta}) + [V\,ies(\hat{\theta})]^2 \quad (2.23)$$

onde o símbolo $E[\cdot]$ denota o valor esperado e $V\,ies(\hat{\theta}) = E[\hat{\theta}] - \theta$ representa o vício.

Note que quanto menor a variância e menor o vício do estimador, melhor será esse modelo e portanto segundo esse critério, o melhor modelo terá o menor valor de EQM.

3. Resultados e Discussão

Dados artificiais são gerados para avaliar o ajuste do modelo de regressão linear. Também é possível avaliar se os critérios de seleção de modelos utilizados conseguem identificar o modelo utilizado na geração ao compará-lo com outros modelos. A Seção contém os resultados para o ajuste do modelo espacial em um conjunto de dados artificiais.

3.1 Modelo de Regressão Linear

Considere o modelo de regressão linear proposto na Subseção 2.2.1 e que possui a seguinte forma

$$Y_i = X_i' \beta + e_i \quad (3.1)$$

em que $e' = (e_1, \ldots, e_n) \sim N_n(0, \sigma^2)$ e $i = 1, \ldots, n$.

Na Subseção 3.1.1 há a descrição sobre o conjunto de dados simulados, o ajuste desse conjunto utilizando diferentes covariáveis conforme o modelo proposto e a avaliação dos critérios de seleção de modelos.

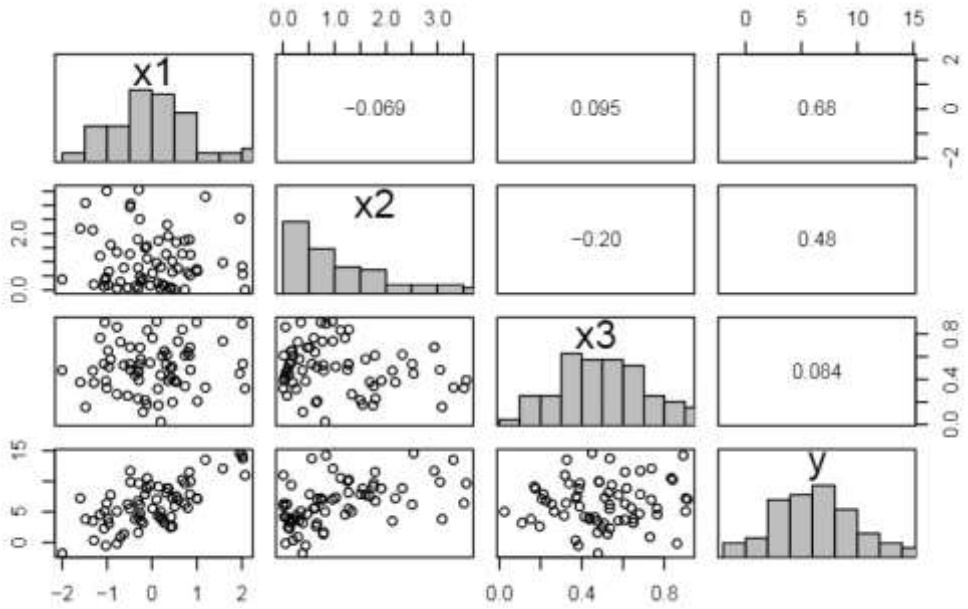
3.1.1 Dados artificiais

Para gerar um conjunto de dados simulados de tamanho n da variável resposta Y_i, $i = 1, \ldots, n$, definida na Equação 3.1, utilizou-se que os seguintes valores: tamanho amostral $n = 67$, vetor de coeficientes $\beta' = (2, 0 ; 3, 0 ; 2, 5 ; 3, 5)$, sendo que o primeiro valor refere-se ao intercepto representado por β_0 e os demais representam os efeitos das covariáveis e a variância $\sigma^2 = 4$.

As variáveis independentes $X_i' = (X_{i0}, X_{i1}, X_{i2}, X_{i3})$ foram geradas de forma independente para cada unidade $i = 1, \ldots, n$ da seguinte forma: $X_{i0} = 1$ para ter intercepto, $X_{i1} \sim N(0, 1)$, $X_{i2} \sim Gama(1, 1)$ e $X_{i3} \sim Beta(2, 2)$. Note que utilizou- se 3 covariáveis e um intercepto.

A Figura 1 apresenta os seguintes gráficos descritivos dos dados: gráficos de dispersão entre a variável resposta e cada uma das covariáveis e entre as covariáveis, histogramas de cada uma dessas variáveis e a correlação entre essas variáveis. Note que a primeira covariável possui uma correlação linear de 0,68 com a variável resposta. Essa é uma forte correlação e por isso tem-se que o gráfico de dispersão entre essas variáveis possui um formato próximo a uma reta. A segunda covariável possui uma correlação ligeiramente menor com a variável resposta (0,48) e portanto percebe-se uma dispersão maior de uma reta. A terceira covariável possui uma correlação muito baixa com a variável resposta (de 0,084) e por isso não é possível identificar um padrão no gráfico de dispersão respectivo. Como as covariáveis foram geradas de forma independentes, os respectivos gráficos de dispersão de x_{k1} versus x_{k2}, sendo k1, k2 = 1, 2, 3 e $x_k = (x_{k1}, \ldots, x_{kn})$, apresentam aleatoriedade e suas correlações também possuem valores muito baixos.

Figura 1: Análise descritiva dos dados. Na diagonal principal há os histogramas das covariáveis x e da variável resposta y. Na matriz acima da diagonal principal há as correlações entre as variáveis e na matriz abaixo dessa diagonal há os gráficos de dispersão. Cada linha e coluna apresentam informações na seguinte ordem: x_1, x_2, x_3 e y.



Fonte: Autor.

Considere que o vetor de parâmetros $\theta = (\beta, \tau)$, sendo $\tau = \sigma^{-2}$, seja desconhecido. Sob o enfoque Bayesiano, considere a priori que β e τ sejam independentes e que $\beta \sim N_4(0, 1000I_4)$ e $\tau \sim$ Gama(2, 1), sendo I_4 uma matriz identidade de ordem 4. Os valores dos hiperparâmetros da distribuição a priori foram escolhidos de forma a obter uma distribuição não informativa para os parâmetros.

Para avaliar a capacidade de a metodologia utilizada recuperar os verdadeiros valores dos parâmetros desconhecidos e de reconhecer o modelo correto, suponha os seguintes modelos:

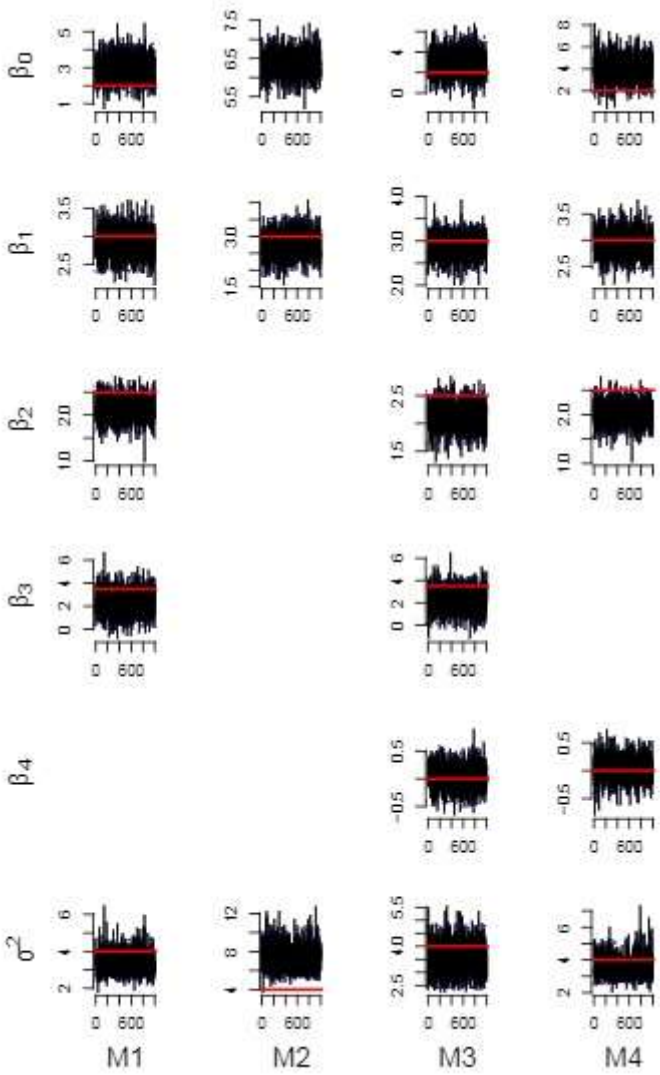
- M1: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + e$;
- M2: $Y = \beta_0 + \beta_1X_1 + e$;
- M3: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + e$;
- M4: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_4X_4 + e$.

Note que o modelo M1 é utilizado para gerar os dados simulados. O modelo M2 retira 2 covariáveis utilizadas na geração dos dados e os modelos M3 e M4 são versões incluindo uma covariável que não foi utilizada na geração dos dados.

Já que a distribuição a posteriori de θ não é conhecida, portanto é necessário recorrer aos métodos de MCMC para obter amostras a posteriori dos parâmetros. Para este caso em particular utiliza-se o Amostrador de Gibbs que requer a obtenção das DCCP dos parâmetros τ e β, conforme descrito na Equação em 2.9. Rodou-se 1100 iterações, descartando as 100 primeiras e portanto obteve-se uma amostra de tamanho 1.000.

Verifica-se na Figura 2 os traços das cadeias dos parâmetros β. Observe que há indícios de convergência pelo fato de manter um padrão aleatório dentro do limite do intervalo de credibilidade 95%, delimitado pelas linhas azuis pontilhadas. Os verdadeiros valores dos parâmetros usados para gerar os dados estão representados pelas linhas vermelhas. Note que esses valores estão contemplados de forma satisfatória nos intervalos de credibilidade dos modelos M1 e M3, conforme esperado uma vez que o M1 é o verdadeiro modelo utilizado na geração dos dados e o modelo M3 difere do M1 por incluir uma variável adicional que não foi utilizada na geração e, portanto pode-se considerar que tenha efeito nulo na variável resposta. Nos demais modelos esses valores não precisavam ser contemplados, pois covariáveis usadas na geração dos dados foram retiradas nos ajustes. Note que não é necessário que o algoritmo faça muitas iterações para que consiga retornar boas estimativas.

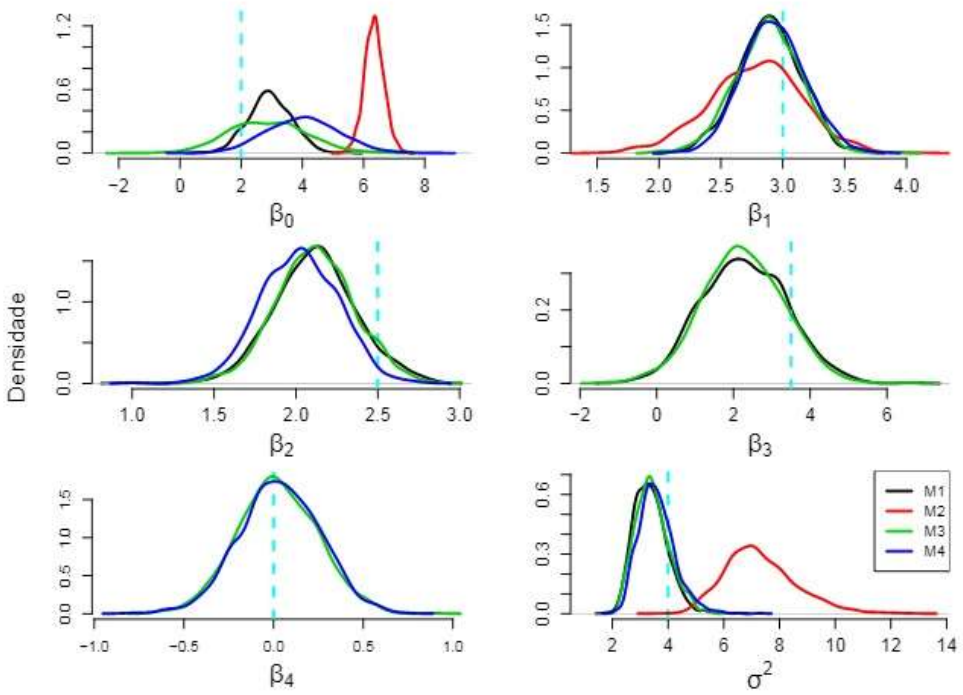
Figura 2: Traço das cadeias e intervalo de credibilidade nas linhas azuis pontilhadas e linha vermelha identificando valor real para os dados artificiais.



Fonte: Autor.

Já na Figura 3 pode ser verificado as densidades dos parâmetros β e $\sigma^2 = 1/\tau$, nos diferentes modelos propostos, referenciados pelas cores, além das linhas tracejadas verticais indicando os valores reais. As linhas pretas referem-se ao M1, as linhas vermelhas ao M2, as linhas verdes ao M3 e as linhas azuis ao M4. Como esperado as densidades possuem comportamentos semelhantes nos modelos M1 e M3. O modelo que mais difere é o M2.

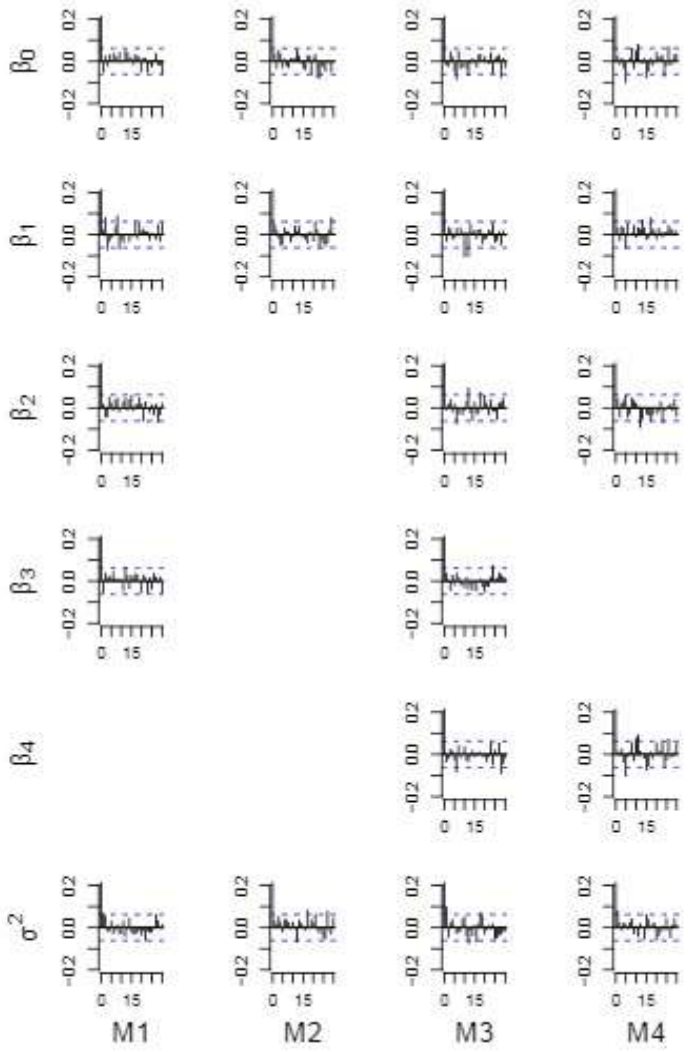
Figura 3: Densidades dos parâmetros β e σ^2 a posteriori e linhas tracejadas referentes aos valores reais dos parâmetros.



Fonte: Autor.

A Figura 4 contém os gráficos de autocorrelação dos parâmetros desconhecidos e pode ser visto que há indícios que os parâmetros β e σ^2 não sejam autocorrelacionados, pois decai rapidamente para dentro do intervalo e por isso não foi necessário utilizar algum espaçamento.

Figura 4: Gráfico de autocorrelação dos parâmetros para os diferentes modelos.



Fonte: Autor.

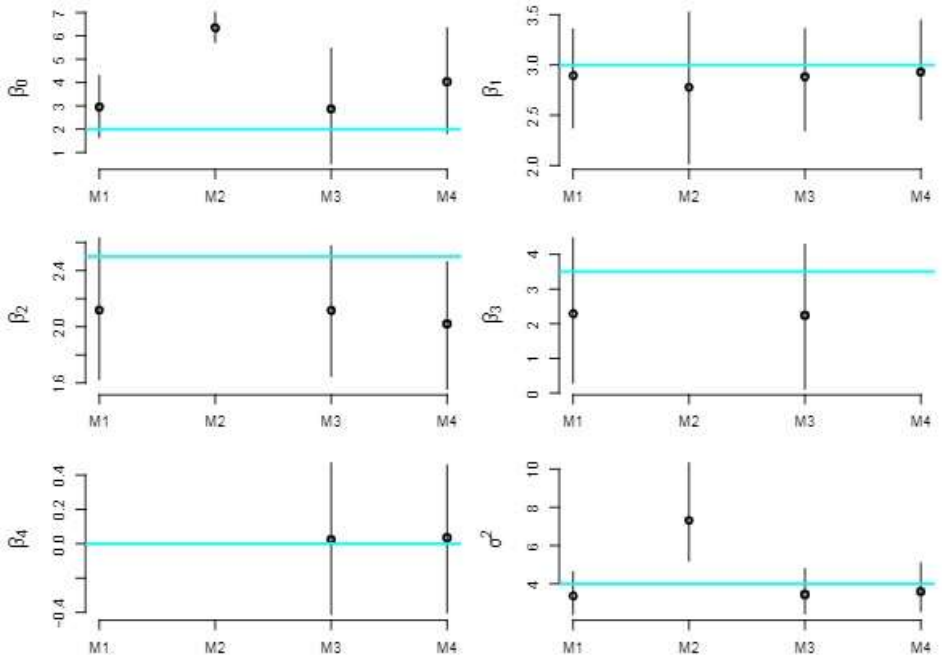
Após verificada as convergências das cadeias, o próximo passo é conferir as estimativas dos parâmetros e analisar as suposições dos resíduos. Na Tabela 1 e na Figura 5 pode-se verificar o resumo das estimativas pontuais e intervalares a posteriori dos parâmetros amostrados comparado aos seus valores reais. Repare que as médias estão bem próximas dos valores reais e que todos esses valores estão incluídos nos seus respectivos intervalos de credibilidade. Observe que os valores amostrados a posteriori da variância no M2 se destoa entre os outros modelose possui maior variabilidade, já que duas variáveis foram retiradas do modelo ao qual gerou-se os dados simulados.

Tabela 1: Coeficientes estimados e seus respectivos intervalos de credibilidade de 95% e os valores reais segundo os modelos propostos.

Modelos	Parâmetro	Médias	$IC_{2,5\%}$	$IC_{97,5\%}$	Valores reais
M1	β_0	2,94	1,65	4,30	2
	β_1	2,89	2,38	3,35	3
	β_2	2,12	1,63	2,63	2,5
	β_3	2,29	0,29	4,47	3,5
	σ^2	3,37	2,41	4,62	4
M2	β_0	6,3	5,7	7,0	2
	β_1	2,8	2,0	3,5	3
	σ^2	7,3	5,2	10,3	4
M3	β_0	2,860	0,531	5,458	2
	β_1	2,880	2,347	3,361	3
	β_2	2,116	1,649	2,576	2,5
	β_3	2,243	0,122	4,290	3,5
	β_4	0,023	-0,411	0,468	
	σ^2	3,443	2,428	4,781	4
M4	β_0	4,026	1,835	6,326	2
	β_1	2,929	2,458	3,441	3
	β_2	2,021	1,559	2,459	2,5
	β_4	0,035	-0,401	0,456	
	σ^2	3,599	2,569	5,068	4

Fonte: Autor.

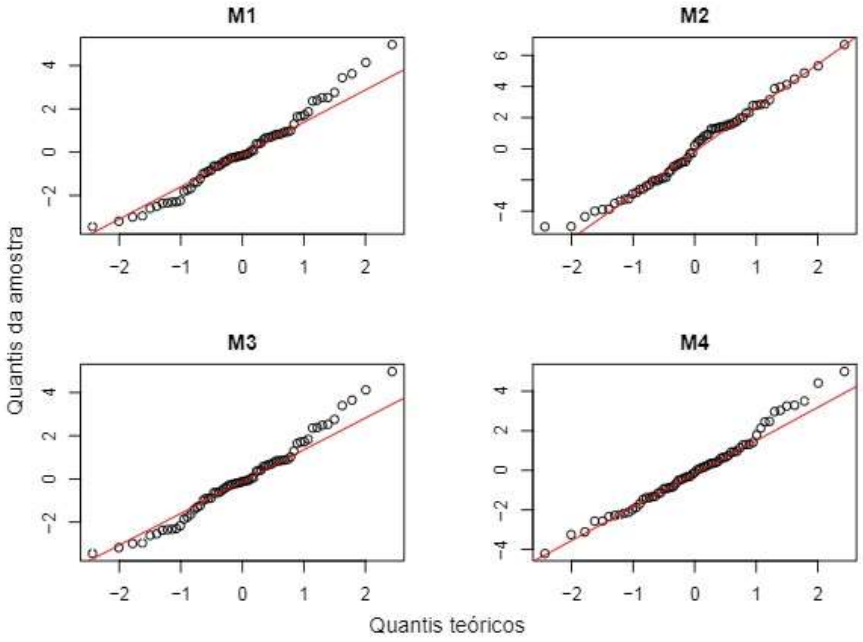
Figura 5: Gráfico de médias a posteriori e intervalos de credibilidade de 95% dos parâmetros β e σ^2. A linha horizontal azul corresponde ao valor verdadeiro.



Fonte: Autor.

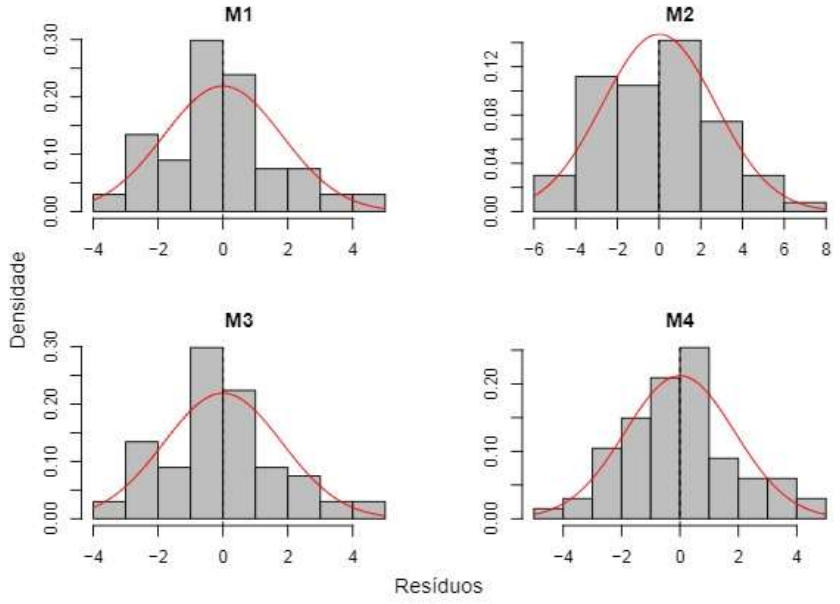
A suposição de normalidade dos resíduos pode ser visualizada a partir do histograma ou de um gráfico q × q que compara os quantis de 2 distribuições e nesse caso compara os quantis dos resíduos com base na amostra a posteriori obtida com os quantis teóricos de uma distribuição normal. No segundo gráfico, considera-se que o modelo foi bem ajustado se os pontos estiverem próximos a uma reta. As Figuras 6 e 7 contêm estes gráficos e, aparentemente, a suposição de normalidade não é violada em nenhum dos modelos propostos, embora haja uma dispersão maior nas caldas.

Figura 6: Gráfico q × q para verificar a suposição de normalidade.



Fonte: Autor.

Figura 7: Histograma dos resíduos e distribuição normal sobreposta ao histograma.



Fonte: Autor.

A Tabela 2 apresenta os p-valores dos testes de Shapiro-Wilk e Kolmogorov-Smirnov para verificar a suposição de normalidade dos resíduos. Comparando esses valores com o nível de significância usual de 5%, não há indícios para rejeitar essa hipótese, conforme suspeitado visualmente na Figura 7.

Tabela 2: Teste de normalidade dos resíduos segundo os modelos propostos

Modelos	p-valor (Shapiro-Wilk)	p-valor (Kolmogorov-Smirnov)
M1	0,36	0,68
M2	0,42	0,71
M3	0,37	0,61
M4	0,54	0,91

Fonte: Autor

A Tabela 3 contém os valores do EQM e DIC para os diferentes modelos. Os modelos M1 e M3 possuem o mesmo valor de EQM, pois ajuste identificou efeito nulo à covariável X_4 e o modelo M1 possui o menor DIC. Sendo assim, considera-se o modelo M1 o melhor modelo comparado aos demais, conforme esperado, já que este modelo gerou os dados.

Tabela 3: Valores de EQM e DIC segundo os modelos propostos.

Modelos	EQM	DIC
M1	3,27	279,67
M2	7,24	328,88
M3	3,27	280,96
M4	3,48	283,67

Fonte: Autor

3.2 Modelo Espacial

Os modelos utilizados na Seção anterior consideraram independência entre as unidades amostrais após eliminar os efeitos das covariáveis na variável resposta. Essa Seção avaliará modelos espaciais que acomodam dependência espacial mesmo após eliminar os efeitos das covariáveis. Os dados foram gerados e ajustados através do programa R Core Team (2019).

Considere o modelo proposto na Seção 2.3 e portanto a variável resposta de cada região com coordenada s_i pode ser escrita da seguinte forma

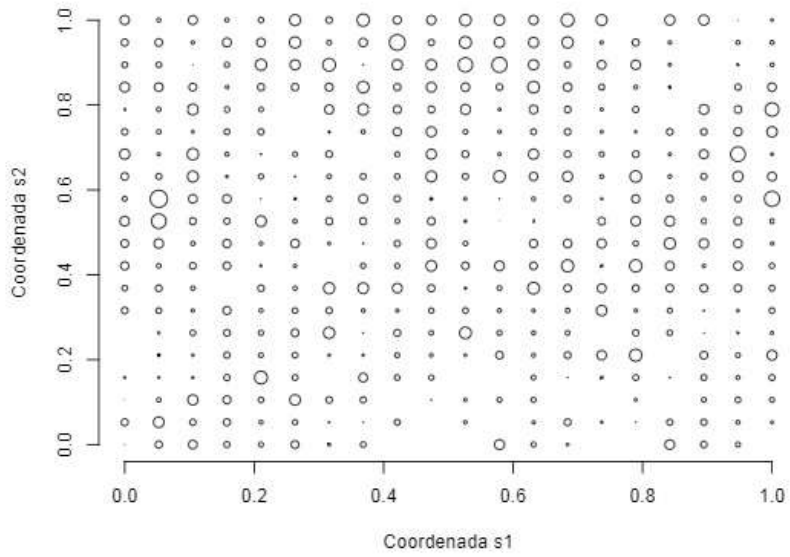
$$Y(s_i) = X(s_i)' \beta + \omega(s_i), \quad (3.2)$$

em que $\omega(s_i)$ tem distribuição normal com média zero, variância V e a correlação entre $\omega(s_i)$ e $\omega(s_j)$ é dada pela função de correlação exponencial, $\exp\{-h \times v\}$. Considere $d_{ij} = h$, para i, j = 1, . . . , n, com d_{ij} medindo a distância euclidiana entre essas regiões.

Para avaliar a capacidade de ajuste desse modelo, dados simulados são gerados. Suponha que cada região possa ser descrita por 2 coordenadas construídas através de uma grade regular dividindo cada eixo unitário em 20 pontos equidistantes

totalizando n = 400 localizações. A Figura 8 apresenta as coordenadas simuladas para cada quantidade da variável Y (s_i) em cada localização específica.

Figura 8: Localizações simuladas.



Fonte: Autor

Assuma os seguintes valores para o vetor de coeficientes: $\beta = (\beta_0 = 2 ; \beta_1 = 1)$. O primeiro valor refere-se ao intercepto e o segundo o efeito da covariável. Para a variância atribua o valor V = 1. Note que o parâmetro de correlação espacial precisa ser positivo, ou seja, $\nu > 0$. Neste trabalho são atribuídos a este parâmetro espacial valores de: 10, 00; 4, 24; 2, 30, representando modelos com baixa, média e alta correlação espacial.

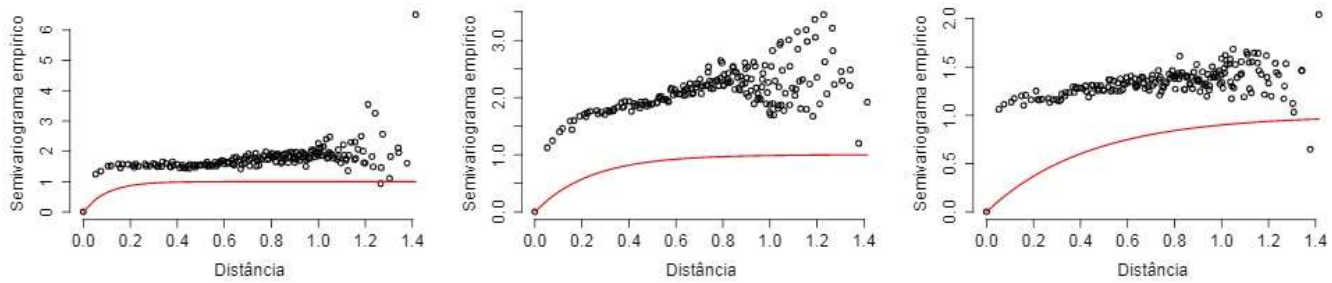
O vetor de covariáveis da região com coordenadas s_i, com i = 1, . . . , n, é dado por $X(s_i) = (X_0(s_i) = 1, X_1(s_i))'$, onde a primeira componente está relacionada ao intercepto e a segunda é gerada da distribuição normal padrão iid, ou seja, $X_1(s_i) \sim N(0, 1)$.

Denote os diferentes conjuntos de dados amostrados da seguinte forma:

- D1: $\nu = 10, 00$;
- D2: $\nu = 4, 24$; e
- D3: $\nu = 2, 30$.

Na Figura 9 apresenta os semivariogramas empíricos dos dados simulados D1, D2 e D3 nos diferentes valores de ν, ou seja, as correlações espaciais dada por $\exp\{-d_{ij} \times \nu\}$, para i, j = 1, . . . , n. Pode-se notar a medida que o valor de ν diminui, as amostras simuladas de Y tendem a apresentar correlação espacial com maior intensidade. Note, também, que o valor do alcance efetivo aumenta na mesma medida.

Figura 9: Semivariograma empírico para os dados simulados D1, D2 e D3 nos diferentes valores de ν.

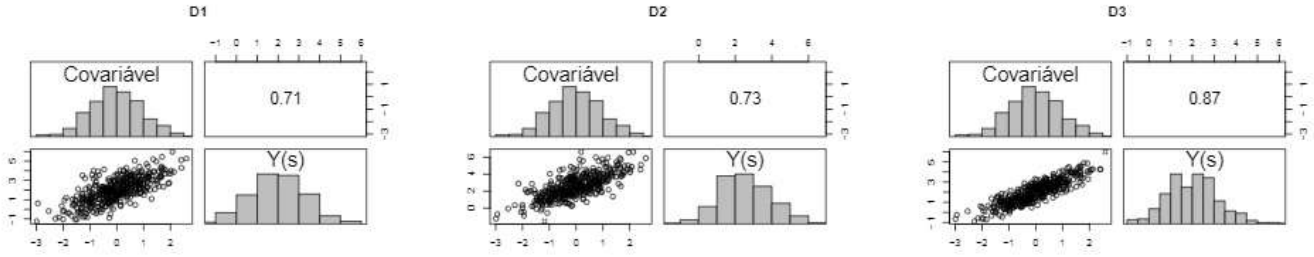


(a) $\nu = 10,00$ (b) $\nu = 4,24$ (c) $\nu = 2,30$

Fonte: Autor

Na Figura 10 há nas diagonais principais os histogramas das variáveis respostas e das covariáveis para cada conjunto de dados gerados. Acima dessas diagonais, há os valores das correlações entre essas variáveis e abaixo há os gráficos de dispersão. Note que em todos os 3 conjuntos de dados utilizados há uma correlação positiva e forte entre a variável resposta e a covariável. Essa forte relação também é percebida pelo formato semelhante de uma reta nos gráficos de dispersão.

Figura 10: Análise descritiva dos diferentes conjuntos de dados. Na diagonal principal há os histogramas da covariável e da variável resposta. Acima da diagonal principal há a correlação entre a variável resposta e a covariável e abaixo dessa diagonal há o gráfico de dispersão.



(a) $\nu = 10,00$ (b) $\nu = 4,24$ (c) $\nu = 2,30$

Fonte: Autor

O vetor de parâmetros desconhecidos sob o modelo espacial é $\theta = (\beta, V, \nu)$.

Para comparar com o modelo espacial dado pela Equação em 3.2, ajustou-se os conjuntos de dados D1, D2 e D3 ao MRLS descrito na Subseção 2.2.1. Sob esse modelo, o vetor de parâmetros desconhecidos é $\theta^{MRL} = (\beta, V = \sigma^2)$.

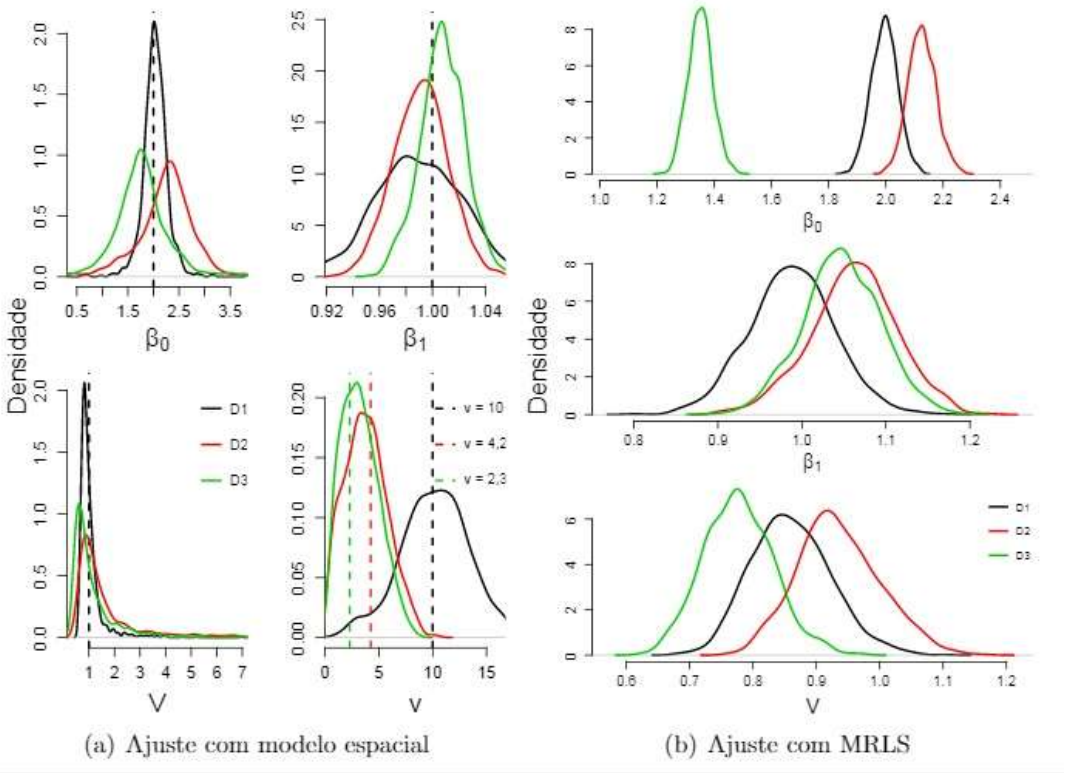
Considere todos os parâmetros independentes a priori e que $\beta_k \sim N(0, 1000)$, $k = 1, 2$, $V^{-1} \sim$ Gama(2, 1) e $\nu \sim$ Gama($d^2_{max}/20$; $d_{max}/10$) . Dessa forma, a priori considera-se que o valor médio de ν é a metade da distância máxima e que tem variância 5.

Para o modelo espacial, rodou-se 51.500 iterações, descartando as 10.500 primeiras e utilizou-se o espaçamento k = 41, obtendo-se então uma amostra de tamanho 1.000, com base nas Figuras 13 e 14. E para o MRLS, rodou-se 1.100 iterações,

descartando as 100 primeiras e utilizou-se o espaçamento k = 1, obtendo-se então uma amostra de tamanho 1.000. Note que o modelo espacial precisou de mais iterações e consequentemente maior tempo computacional para obter convergência.

A Figura 11 apresenta as densidades a posteriori dos parâmetros sob o modelo espacial e sob o MRLS. As linhas pretas tracejadas indicam os valores reais utilizados para estimar a cadeia. Note que sob o modelo espacial, as amostras a posteriori dos parâmetros nos 3 conjuntos de dados parecem estar em torno dos valores verdadeiros desses parâmetros.

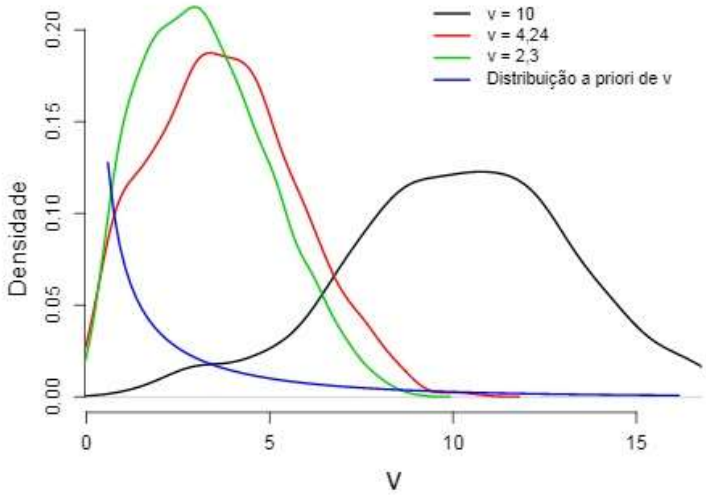
Figura 11: Densidades a posteriori dos parâmetros amostrados nos diferentes tipos de ajustes.



Fonte: Autor

Na Figura 12, encontra-se a distribuição a priori de ν em conjunto com as densidades das amostras a posteriori de ν = 10, 00, ν = 4, 24 e ν = 2, 30. Note que as densidades das amostras a posteriori de ν diferem-se da distribuição a priori de ν, indicando que os valores observados da variável resposta trouxeram novas informações sobre esse parâmetro.

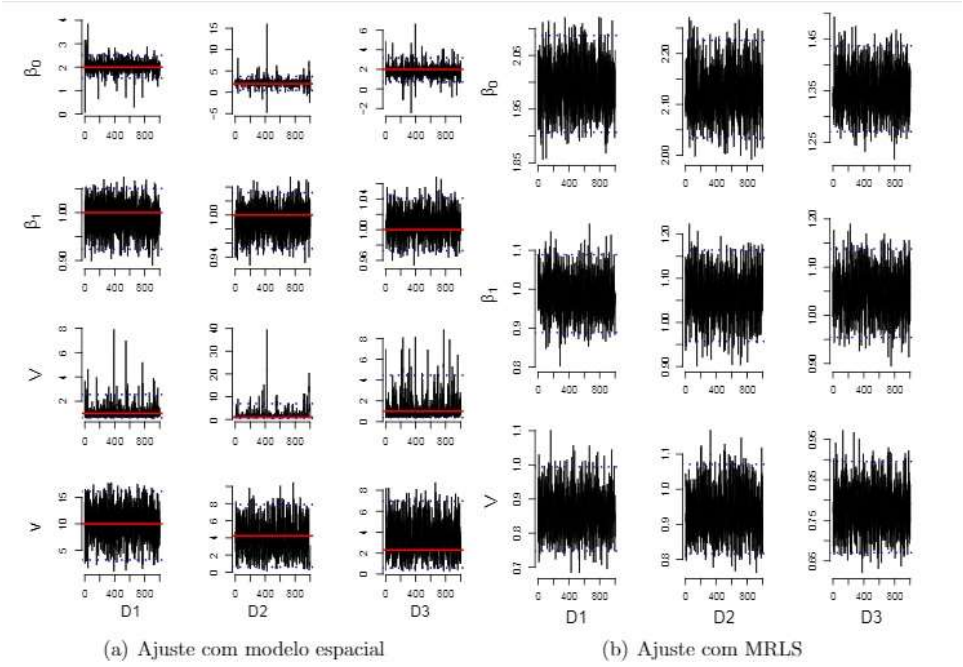
Figura 12: Distribuição a priori de ν em conjunto com as densidades das amostras a posteriori de ν = 10, ν = 4, 24 e ν = 2, 3.



Fonte: Autor

Já a Figura 13 apresenta os traços das cadeias dos parâmetros amostrados. Observe que há indícios de convergência pelo fato de manter um padrão aleatório dentro do limite do intervalo de credibilidade de 95%, delimitado pelas linhas azuis pontilhadas. Os valores reais estão representados pelas linhas vermelhas. Note que todos os valores reais estão contemplados nos intervalos de credibilidade.

Figura 13: Traço das cadeias, intervalo de credibilidade de 95% nas linhas azuis pontilhadas e valores reais nas linhas vermelhas sob diferentes tipos de ajustes.

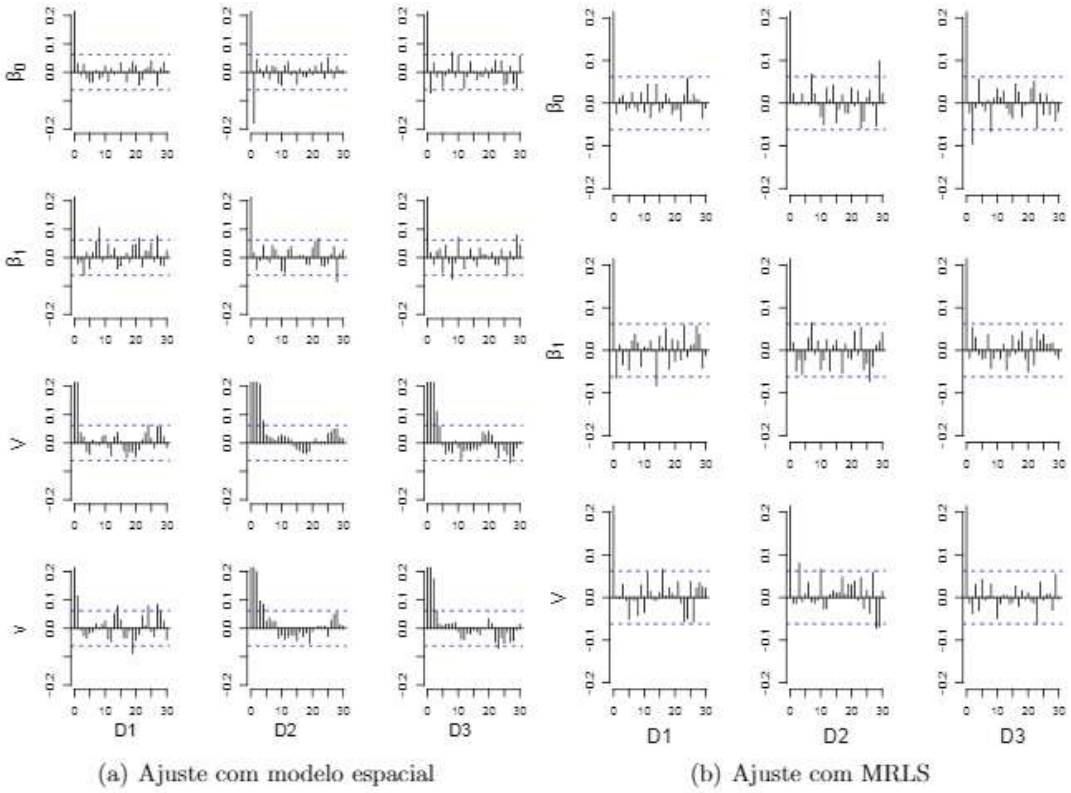


(a) Ajuste com modelo espacial (b) Ajuste com MRLS

Fonte: Autor

A Figura 14 apresenta os gráficos de autocorrelação destes parâmetros e pode ser visto que há indícios de que os parâmetros β e V não são autocorrelacionados, pois decai rapidamente para dentro do intervalo.

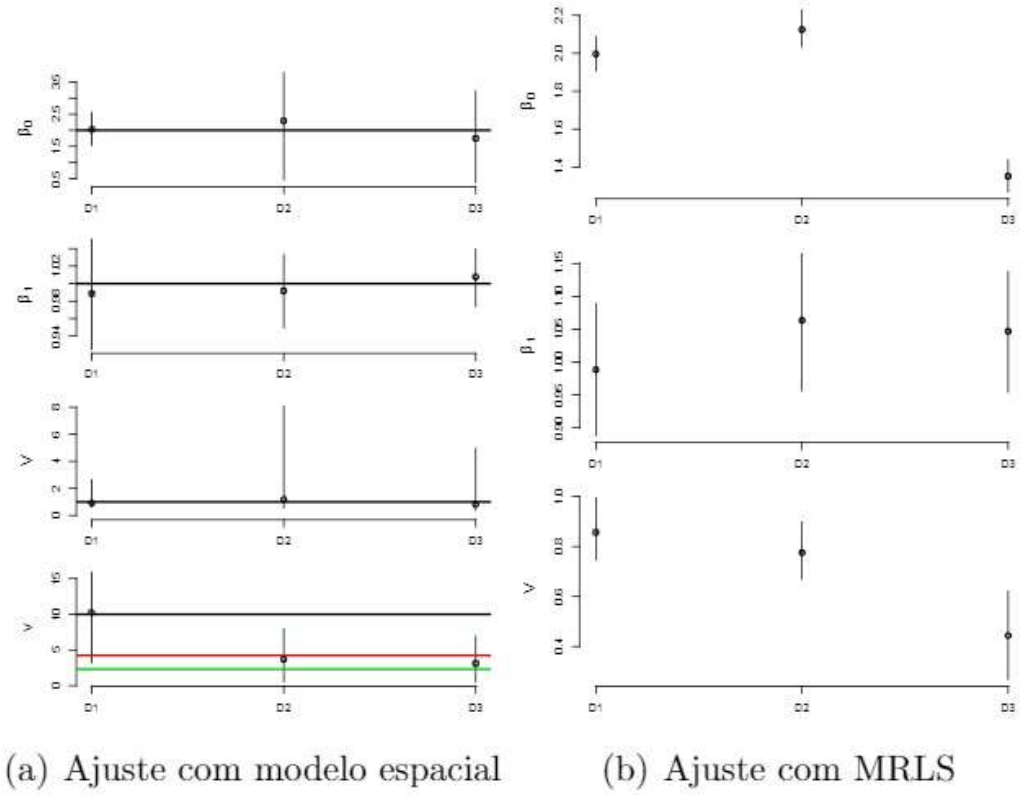
Figura 14: Gráfico de autocorrelação dos parâmetros sob diferentes tipos de ajustes.



Fonte: Autor

A Figura 15 e a Tabela 4 contêm o resumo das estimativas a posteriori pontuais e intervalares dos parâmetros amostrados sob diferentes tipos de ajustes comparados aos seus valores reais. Os parâmetros destacados com $*$ referem-se à mediana a posteriori dos parâmetros desconhecidos ajustados com MRLS. Repare que os valores estão bem próximos dos valores reais no caso do modelo espacial (ME) e que todos estão incluídos nos seus respectivos intervalos de credibilidade de 95%. Como os dados foram gerados do modelo espacial, o ajuste do MRLS ou não consegue recuperar o valor verdadeiro dos parâmetros ou possui uma alta variabilidade.

Figura 15: Média e intervalo de credibilidade a posteriori para os diferentes conjuntos de dados amostrados de Y (s) e diferentes tipos de ajustes.



Fonte: Autor

A Figura 16 apresenta os gráficos de dispersão dos resíduos versus as unidades e dos resíduos versus as covariáveis sob os diferentes conjuntos de dados e os 2 modelos ajustados. Para o conjunto de dados D1 tem-se que os resíduos possuem comportamento aleatório e em torno de zero sob ambos os modelos. Além disso, eles possuem comportamentos semelhantes sob os 2 modelos ajustados. Esse resultado era esperado uma vez que o primeiro conjunto de dados foi gerado com um alto valor de ν acarretando em um modelo com fraca dependência espacial. Porém, há de se observar que nas outras amostras modeladas pelo MRLS, os resíduos se comportam como curvas, sugerindo que o MRLS não obteve um bom ajuste. Sob o modelo espacial, os resíduos parecem ter um comportamento aleatório em torno do zero.

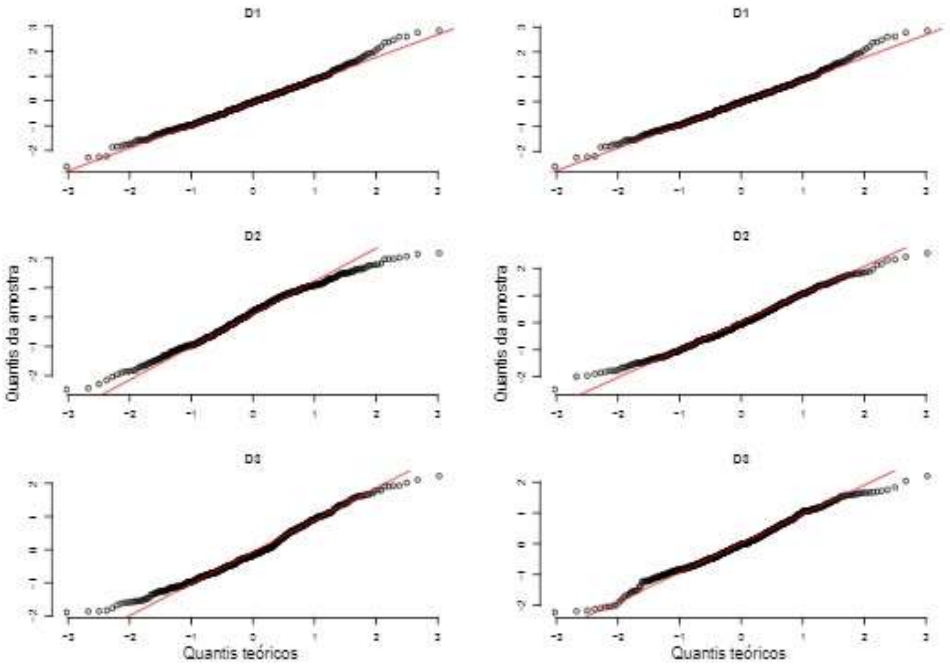
Figura 16: Distribuição dos resíduos em torno do zero sob diferentes tipos de ajustes.

Dados	Parâmetro	Mediana *a posteriori*	$IC_{2,5\%}$	$IC_{97,5\%}$	Valores reais
D1	β_0	2,03	1,53	2,51	2
	β_0^*	2,00	1,91	2,09	-
	β_1	0,99	0,92	1,05	1
	β_1^*	0,99	0,89	1,09	-
	V	0,91	0,64	2,55	1
	V^*	0,86	0,75	0,99	-
	ν	10,23	3,17	16,16	10
D2	β_0	2,29	0,44	3,71	2
	β_0^*	2,12	2,03	2,23	-
	β_1	0,99	0,95	1,03	1
	β_1^*	1,06	0,96	1,16	-
	V	1,17	0,58	6,95	1
	V^*	0,93	0,82	1,07	-
	ν	3,69	0,59	7,89	4,24
D3	β_0	1,75	0,72	3,17	2
	β_0^*	1,35	1,27	1,44	-
	β_1	1,01	0,97	1,04	1
	β_1^*	1,05	0,95	1,14	-
	V	0,82	0,40	4,46	1
	V^*	0,78	0,67	0,90	-
	ν	3,12	0,59	6,96	2,30

Fonte: Autor

As Figuras 17 e 18 apresentam o histograma e o gráfico q × q nas diferentes amostras de Y (s) e diferentes tipos de ajuste. Aparentemente, a suposição de normalidade não é violada, já que os pontos se dispôs o longo da reta em vermelho, embora haja valores atípicos com caldas pouco pesadas em ambos tipos de ajustes.

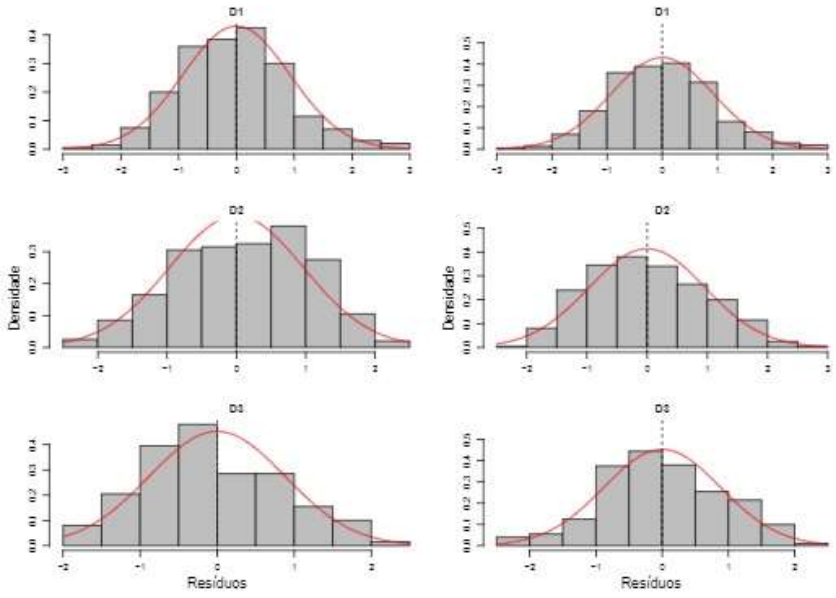
Figura 17: Gráfico q × q para verificar a suposição de normalidade sob diferentes tipos de ajustes.



(a) Ajuste com modelo espacial (b) Ajuste com MRLS

Fonte: Autor

Figura 18: Histograma dos resíduos e distribuição normal sobreposta ao histograma sob diferentes tipos de ajustes.

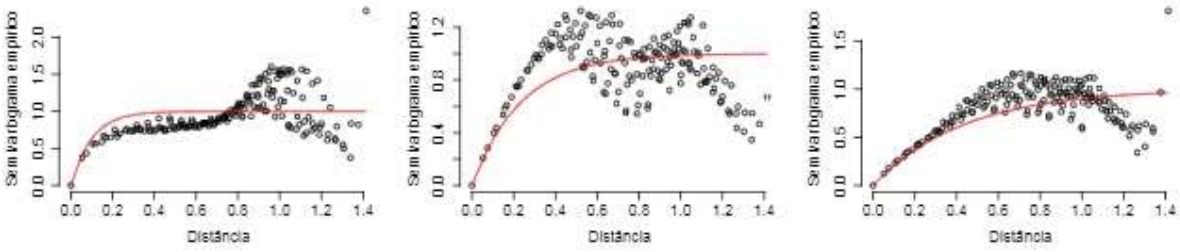


(a) Ajuste com modelo espacial (b) Ajuste com MRLS

Fonte: Autor

Na Figura 19 apresenta os semivariogramas empíricos dos resíduos nos diferentes valores de ν.

Figura 19: Semivariograma empírico dos resíduos nos diferentes valores de ν.



(a) $\nu = 10,00$ (b) $\nu = 4,24$ (c) $\nu = 2,30$

Fonte: Autor

Tabela 4: Parâmetros ajustados do modelo para os diferentes conjuntos de dados amostrados e diferentes tipos de ajustes.

Dados	Parâmetro	Mediana *a posteriori*	$IC_{2,5\%}$	$IC_{97,5\%}$	Valores reais
D1	β_0	2,03	1,53	2,51	2
	β_0^*	2,00	1,91	2,09	-
	β_1	0,99	0,92	1,05	1
	β_1^*	0,99	0,89	1,09	-
	V	0,91	0,64	2,55	1
	V^*	0,86	0,75	0,99	-
	ν	10,23	3,17	16,16	10
D2	β_0	2,29	0,44	3,71	2
	β_0^*	2,12	2,03	2,23	-
	β_1	0,99	0,95	1,03	1
	β_1^*	1,06	0,96	1,16	-
	V	1,17	0,58	6,95	1
	V^*	0,93	0,82	1,07	-
	ν	3,69	0,59	7,89	4,24
D3	β_0	1,75	0,72	3,17	2
	β_0^*	1,35	1,27	1,44	-
	β_1	1,01	0,97	1,04	1
	β_1^*	1,05	0,95	1,14	-
	V	0,82	0,40	4,46	1
	V^*	0,78	0,67	0,90	-
	ν	3,12	0,59	6,96	2,30

Fonte: Autor

A partir da Tabela 5, verifica-se os p-valores dos testes de Shapiro-Wilk (S-W) e Kolmogorov-Smirnov (K-S), além dos EQM e DIC para as diferentes amostras de Y (s) e diferentes tipos de ajustes. Note que, ao nível de significância de 5%, os p-valores do teste de Kolmogorov-Smirnov não rejeitaram as hipóteses de normalidade dos MRLS em todos os 3 conjuntos de dados enquanto os do Shapiro-Wilk só não rejeitaram no conjunto de dados D1. Para o modelo espacial, todos os testes rejeitaram a hipótese de normalidade exceto para o conjunto de dados D1, que era esperado. Acredita-se que essa rejeição no modelo espacial deva-se ao fato dos resíduos terem dependência espacial.

Tabela 5: Teste de normalidade dos resíduos e taxa de aceitação (TA) segundo as amostras de Y (s) a partir dos tipos de ajuste.

Dados	Tipo de ajuste	p-valor (S-W)	p-valor (K-S)	TA
D1	MRLS	0,058	0,90	
	ME	0,05	0,14	0,64
D2	MRLS	0,010	0,54	
	ME	0,0006	0,0005	0,66
D3	MRLS	0,009	0,44	
	ME	9,42e-05	6,44e-05	0,66

Fonte: Autor

Segundo o DIC, na Tabela 6, o modelo espacial é considerado o melhor modelo em todos os 3 conjuntos de dados. O EQM não consegue diferenciar o modelo espacial do MRLS.

Tabela 6: Valores dos EQM e DIC segundo as amostras de Y (s) a partir dos tipos de ajuste.

Dados	Tipo de ajuste	EQM	DIC
D1	MRLS	0,85	1.077,85
	ME	0,85	851,31
D2	MRLS	0,92	1.111,19
	ME	0,94	588,53
D3	MRLS	0,77	1.037,36
	ME	0,77	395,01

Fonte: Autor

4. Conclusão

Primeiramente, verificou-se a capacidade de estimação do modelo de regressão linear sob o enfoque Bayesiano utilizando dados simulados. Analisou-se 5 modelos distintos em que retirou-se e incluiu-se variáveis que pertenciam e que não pertenciam ao modelo que originou os dados simulados. Para obter amostras das distribuições a posteriori destes parâmetros, foi necessário implementar o Amostrador de Gibbs e obter as cadeias necessárias para estimação. Através dos gráficos dos traços das cadeias e de autocorrelação, constatou-se convergência do método utilizado com poucas iterações. Não foi necessário utilizar espaçamento e utilizou-se um baixo período de aquecimento. Quando ajustou-se o modelo utilizado para gerar os dados, os verdadeiros valores dos parâmetros estavam dentro do intervalo de credibilidade, reforçando a eficiência do ajuste do modelo. Além disso, através da comparação entre modelos, o EQM e o DIC selecionaram o modelo verdadeiro mostrando a eficiência do método.

O modelo de regressão linear supõe independência entre as unidades após eliminar o efeito das covariáveis. Essa suposição é bastante forte e difícil de ser obtida na prática. Sendo assim, analisou-se um modelo espacial com função de covariância exponencial para modelar dados geoestatísticos. Para verificar a capacidade de inferir sobre os parâmetros desconhecidos, gerou-se 3 conjuntos de dados simulados e estimou-se os parâmetros. Os valores verdadeiros foram recuperados de forma satisfatória demonstrando a eficiência do ajuste. Ajustou-se um modelo de regressão linear simples aos dados gerados e comparou-se com o modelo espacial. O modelo espacial precisou de mais iterações do MCMC para obter convergência e com isso um maior tempo computacional. Os resíduos do MRLS pareceram não serem independentes sob 2 conjuntos de dados e no outro conjunto de dados obteve um comportamento diferente devido a ter uma fraca correlação espacial. O DIC conseguiu identificar o modelo espacial como o melhor modelo para os 3 conjuntos de dados e o EQM não conseguiu diferenciar os modelos.

Para trabalhos futuros, sugiro introduzir conjunto de dados reais quando se tratar de dados com uma ligeira estrutura de correlação espacial, já que foi observado que o ajuste do modelo e modelo espacial comportou-se de forma satisfatória, além da inferência bayesiana. Ademais, o parâmetro espacial não é trivial de ser calculado, portanto pode ser interessante introduzir novas distribuições de probabilidade para esse parâmetro a priori como alternativa a diminuir sua variabilidade ou, até mesmo, alterar os hiperparâmetros.

Referências

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*(1st ed.). Wiley.

Borgoni, R., & Billari, F. C. (2003). Bayesian spatial analysis of demographic survey data: An application to contraceptive use at first sexual intercourse. *Demographic Research*,8, 61-92.

Casella, G., & Berger, R. (2002). *Statistical inference*. Thomson Learning.

Cressie, N. (1993). *Statistics for spatial data*. J. Wiley.

Câmara, G., & Ortiz, M. (1998). Sistemas de informações geográficas para aplicações ambientais e cadastrais: uma visão geral.In: *CONGRESSOBRASILEIRO DE ENGENHARIA AGRÍCOLA*.,27, 59-82.

Dawn B. Woodard, R. L. W., & O'Connell, M. A. (2010). Spatial inference of nitrate concentrations in groundwater. *Journal of Agricultural, Biological, and Environmental Statistics*,15(2), 209-227.

Ehlers, R. S. (2003). Introdução a inferência bayesiana [Computer software manual]. Curitiba. Retrieved from<http://www.leg.ufpr.br/~paulojus/CE227/ce227.pdf>

Gamerman, D. (2004). *Markov chain monte carlo: Stochastic simulation for bayesian inference* (1st ed.). Chapman and Hall/CRC.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain monte carlo: Stochastics imulation for bayesian inference* (2nd ed.). Chapman and Hall/CRC.

Gelfand, A. E., & Smith, A. F. M. (1990). amping-based approaches to calculating marginal densities. *Journal of the American Statistical Association*,85(410),398–409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, andthe bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*,57(6), 721–741.

Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*,57, 97–109.

Krige, D. (1951). A statistical approaches to some basic mine valuation problemson the witwatersrand. *Journal of the Chemical, Metallurgical and Mining Societyof South Africa*,52, 119-139.

Macera, M. H. C. (2011). *Uso dos métodos clássico e bayesiano para os modelosnão-lineares heterocedásticos simétricos* (Ciências de Computação e MatemáticaComputacional). Universidade de São Paulo, São Carlos.

Metropolis, N. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*,21(6), 1087–1092.

Migon, H., Gamerman, D., & Louzada, F. (2014). *Statistical inference: An integrated approach, second edition*. Taylor & Francis.

Morettin, P. A., & de O. Bussab, W. (2010). *Estatistica básica*(3rd ed.). Editora Saraiva.

O'Hagan, A., & Kendall. (1994). *Bayesian inference*(1st ed.). Edward Arnold.

Paulino, C. D., Amaral Turkman, M. A., Murteira, B., & Silva, G. L. (2018). *Estatística bayesiana*(2a ed.). Lisboa: Fundação Calouste Gulbenkian.

Pereira A.S. et al. (2018). *Metodologia da pesquisa científica*. [e-book]. Santa Maria. Ed. UAB/NTE/UFSM.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from<https://www.R-project.org/>

Robert, C. P., & Casella, G. (2004). *Monte carlo statistical methods*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*,7, 110-120.

Roberts, G. O., & Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*,16, 351-367.

Smith, R. L. (1996). *Estimating nonstationary spatial correlations.* Relatório técnico, Cambridge University, UK.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & der Linde, A. V. (2002). Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc*.(64),583-639.

Weisberg, S. (2014). *Applied linear regression*(4th ed.). Hoboken NJ: Wiley.