UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

WANESSA WERIDIANA DA LUZ FREITAS

Análise Exploratória e Regressão Espacial para Dados Simbólicos Intervalares

Recife

2022

WANESSA WERIDIANA DA LUZ FREITAS

Análise Exploratória e Regressão Espacial para Dados Simbólicos Intervalares

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutora em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientadora: Prof. Dra. Renata Maria Cardoso Rodrigues de Souza
Coorientador: Prof. Dr. Getúlio José Amorim do Amaral

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

F866a Freitas, Wanessa Weridiana da Luz
Análise exploratória e regressão espacial para dados simbólicos intervalares / Wanessa Weridiana da Luz Freitas. – 2022.
117 f.: il., fig., tab.

Orientadora: Renata Maria Cardoso Rodrigues de Souza.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.
Inclui referências e apêndices.

1. Inteligência computacional. 2. Modelos de regressão. 3. Dados intervalares. I. Souza, Renata Maria Cardoso Rodrigues de (orientadora). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2023-17

Wanessa Weridiana da Luz Freitas

“Análise Exploratória e Regressão Espacial para Dados Simbólicos Intervalares”

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 15/08/2022.

_____

Orientadora: Profa. Dra. Renata Maria Cardoso Rodrigues de Souza

BANCA EXAMINADORA

_____

Prof. Dr. Nivan Roberto Ferreira Junior
Centro de Informática / UFPE

_____

Prof. Dr. Ronei Marcos de Moraes
Departamento de Estatística / UFPB

_____

Profa. Dra. Roberta Andrade de Araújo Fagundes
Escola Politécnica de Pernambuco / UPE

_____

Profa. Dra. Fernanda de Bastiani
Departamento de Estatística / UFPE

_____

Prof. Dr. Marcus Costa de Araújo
Departamento de Engenharia Mecânica / UFPE

Dedico este trabalho à minha família, em especial a Vanderléa, Márcio e Matteo, que são os pilares da minha força para vencer na vida.

AGRADECIMENTOS

Primeiramente a Deus, por todas as graças a mim concedidas.

Aos meus orientadores Renata Souza e Getúlio José pela excelente orientação nesta tese, pela paciência e dedicação.

A minha mãe, pela confiança, paciência e amor incondicional.

A Márcio de Moura, pelo amor, dedicação, companheirismo, paciência e por sempre acreditar em mim.

Aos familiares, pelo suporte oferecido, em especial a minha avó Luzia e minha tia Verônica.

As amigas que contribuíram para a realização desse trabalho. Em especial a Terezinha, por todos os momentos que passamos juntas. A Luana, que esteve ao meu lado desde a graduação. A Silvany, minha psicóloga preferida. E a "Lauret", por todo carinho.

Aos meus amigos Dácio, Bruno Maciel, Filipe "Karnal", Ozonias, Paulo Henrique, Brunno Wagner, Silas, Sidartha e Halisson que fazem parte da família APG-CIn, em que compartilhamos diversos momentos ao longo desse período. Aos amigos Jodavid, Carlos e Pedro que sempre estiveram na torcida.

Aos colegas do grupo de pesquisa, em especial a Rafaela, Dailys, Flávia, Ieda, Telmo, Victor e Wagner, pelos questionamentos feitos ao longo das reuniões de grupo.

Aos professores do Centro de Informática da Universidade Federal de Pernambuco que contribuíram para minha formação.

A FACEPE, pelo apoio financeiro.

"Sempre seja verdadeiro consigo mesmo e nunca deixe o que outra pessoa diz o distrair de seus objetivos." (OBAMA, 2018).

RESUMO

A Análise de Dados Espaciais é uma área que busca identificar padrões existentes em uma determinada região através de diversas metodologias, como índices e testes de autocorrelação espacial. A compreensão da dependência espacial de um fenômeno em uma dada região pode ser mensurada através de abordagens envolvendo a noção de autocorrelação espacial. Essas técnicas são usualmente empregadas no contexto de variáveis clássicas (pontuais). Por outro lado, a Análise de Dados Simbólicos é uma área de pesquisa e aplicação relacionada às áreas de aprendizagem de máquina e estatística, que fornecem ferramentas para descrever unidades (objetos), permitindo lidar com diversos tipos de variáveis, inclusive variáveis do tipo intervalar. Uma questão relevante consiste em tentar obter uma descrição da autocorrelação espacial para variáveis do tipo intervalar. Assim, neste trabalho buscamos conciliar a análise de dados simbólicos (variáveis intervalares) com a análise de dados espaciais. Mais especificamente, o presente estudo tem o intuito de identificar o comportamento de informações georreferenciadas para dados intervalares na Análise de Dados Simbólicos. Os objetivos principais são: i) estender o índice de autocorrelação espacial de Moran da Análise Exploratória Espacial para o caso de dados intervalares e ii) modelar dados utilizando diferentes modelos de regressão. Para analisarmos os índices de autocorrelação espacial intervalar propostos, realizamos experimentos com conjuntos de dados sintéticos do tipo intervalar. Além disso, analisamos duas aplicações para dados reais. A primeira utiliza dados de notificações de casos de COVID-19 para o nordeste brasileiro e a segunda está relacionada ao preço de aluguel de imóveis na cidade de Munique. No contexto de regressão, utilizamos o modelo de regressão linear para dados do tipo intervalar e um modelo de regressão espacial, que leva em consideração a conectividade existente entre as regiões. Para avaliarmos o desempenho dessas abordagens, realizamos simulações de Monte Carlo em que calculamos a média e o desvio padrão da magnitude média relativa do erro da estimativa dos modelos analisados. Além disso, avaliamos o desempenho dos modelos de regressão em conjuntos de dados reais.

Palavras-chaves: análise de dados simbólicos; análise espacial; dados intervalares; índice de Moran; modelos de regressão.

ABSTRACT

Spatial Data Analysis is a research area that seeks to identify the existence of patterns in a given spatial region by employing appropriate methodologies. The understanding of the spatial dependence of a phenomenon of interest in a given region can be measured through the concept of spatial autocorrelation. Usually, such machinery is employed only in the context of classic variables (punctual variables). On the other hand, Symbolic Data Analysis is a research field related to the framework of machine learning and statistics, which provide us tools to deal with several types of variables, including interval-type variables. A relevant issue regards the description of the spatial autocorrelation in the case of interval type variables. In this work, we seek to unify concepts of symbolic data analysis (for interval variables) and spatial data analysis. More specifically, the present study aims to identify the behavior of georeferenced information for interval data in Symbolic Data Analysis. The main objectives are i) to extend the Moran's spatial autocorrelation index of Spatial Exploratory Analysis for the case of interval data and ii) to investigate such data through regression models. To analyze the proposed interval spatial autocorrelation indexes, we carried out experiments with interval-type synthetic data sets. Also, we have analyzed two applications for real data. The first one considers data involving the number of COVID-19 notifications in northeastern Brazil, while the second application refers to the rental price of houses in the city of *Munich*. In the context of regression, we use the linear regression model for interval-type data and a spatial regression model, which takes into account the existing connectivity between regions. To examine the performance of these approaches, we make Monte Carlo simulations and calculate the mean and standard deviation of the performance metric for the models analyzed.

Keywords: interval data; Moran's index; regression models; symbolic data analysis; spatial analysis.

LISTA DE FIGURAS

Figura 1 – Objeto de pesquisa da tese. 20

Figura 2 – Explanação de uma tabela de dados em que (X, Y) representam variáveis clássicas, para uma tabela de dados simbólicos (X', Y') em que X' e Y' descrevem um conjunto de variáveis simbólicas. 26

Figura 3 – Mapa de pontos de Snow das mortes por cólera em 1854, Golden Square. . 41

Figura 4 – Exemplo de eventos padrões com a localização de monitoramento da poluição do ar em três estados dos EUA. 42

Figura 5 – Exemplo de eventos contínuos com a intensidade da violência por causas externas em Porto Alegre. 43

Figura 6 – Exemplo com taxas agregadas do risco por morte súbita infantil para o período de 1974-1978, na Carolina do Norte. 43

Figura 7 – Representação de um mapa com dados padrões agregados para um mapa com dados simbólicos. 50

Figura 8 – Mapa da cidade de Munique com os centroides por regiões. 54

Figura 9 – Dados simulados com diferentes critérios para a geração da matriz de vizinhança: 1-*nn* (a); 2-*nn* (b); distância Chebyshev (c) e contiguidade da rainha (d). 57

Figura 10 – Valores para os índices utilizando o critério de vizinhança: 1-*nc* (a), 2-*nc* (b), distância de Chebyshev (c) e contiguidade da rainha (d), baseado na matriz **W** com variabilidade no geo-objeto baixa. 67

Figura 11 – Valores para os índices utilizando o critério de vizinhança: 1-*nc* (a), 2-*nc*(b), distância de Chebyshev (c) e contiguidade da rainha (d), baseados na matriz **W** com variabilidade no geo-objeto moderada. 68

Figura 12 – Valores para os índices utilizando os critérios de vizinhança: 1-*nc* (a), 2-*nc* (b), distância de Chebyshev (c) e contiguidade da rainha (d), baseado na matriz **W** com variabilidade no geo-objeto alta. 69

Figura 13 – Valores para o $IASi$, $IASi^*$ e $IASi^*$(amplitude=0) com a matriz $\mathbf{\Lambda}$, utilizando $\mathbf{1}$-*nn* com as distâncias: Gowda-Diday, City block, Euclideana e Hausdorff em (a), (c), (e) e (g), respectivamente; $\mathbf{2}$-*nc* com as distâncias de Gowda-Diday, City block, Euclideana e Hausdorff em (b), (d), (f) e (h), respectivamente. 70
Figura 14 – Média (a) e Amplitude (b) dos preços de aluguéis na cidade de Munique. . 76
Figura 15 – Cidade de Munique dividida em dois grupos(a), e em quatro grupos (b). . . 78
Figura 16 – Mapa do Brasil com destaque para a região Nordeste. 80
Figura 17 – Distribuição espacial da taxa de incidência de dados clássicos (a) e taxa de incidência obtida a partir de dados intervalares (b), para COVID-19 nas microrregiões do Nordeste do Brasil. 81
Figura 18 – Dados intervalares da COVID-19 para os estados: Ceará, Maranhão e Pernambuco. 82
Figura 19 – Distribuição espacial dos municípios considerando o número de óbitos infantis (a), saneamento básico (b), abastecimento de água (c) e coleta de lixo urbano (d). 99
Figura 20 – Distribuição espacial por microrregiões para as variáveis número de óbitos infantis (a), saneamento básico (b), abastecimento de água (c) e coleta de lixo urbano (d). 100
Figura 21 – Histogramas para variáveis intervalares: número de óbitos infantis (a); coleta de lixo urbano (b); taxa de abastecimento de água (c); taxa de saneamento básico (d), referente ao Estado de Pernambuco. 101
Figura 22 – Distribuição espacial dos municípios considerando as variáveis renda per capita (a), percentual das famílias que ganham meio salário mínimo (b), taxa de analfabetismo (c) e densidade demográfica (d). 103
Figura 23 – Distribuição espacial para as micorregiões das variáveis renda per capita (a), percentual de famílias que ganham meio salário mínimo (b), taxa de analfabetismo (c) e densidade demográfica (d). 105
Figura 24 – Histogramas para variáveis intervalares do Nordeste brasileiro: renda per capita (a); percentual de famílias que ganham meio salário mínimo (b), taxa de analfabetismo (c); densidade demográfica (d). 106
Figura 25 – Publicação resultante dessa tese. 117

LISTA DE TABELAS

Tabela 1 – Exemplo de uma tabela de dados simbólicos com os times da copa francesa. 26
Tabela 2 – Conjunto de dados clássicos com gastos (em dólares) do cartão de crédito. 28
Tabela 3 – Conjunto de dados intervalares com gastos do cartão de crédito. 29
Tabela 4 – Exemplo de uma tabela simbólica com dados 'naturalmente' intervalares. . 29
Tabela 5 – Dados veterinários. 32
Tabela 6 – Intervalos de confiança com 5% de erro para os índices de autocorrelação espacial, utilizando a matriz de pesos **W**. 59
Tabela 7 – Índices de autocorrelação espacial para dados clássicos e intervalar, utilizando a matriz de pesos $\mathbf{\Lambda}$. 60
Tabela 8 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com baixa concentração de pontos, utilizando a matriz **W**. 63
Tabela 9 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com média concentração de pontos, utilizando a matriz **W**. 64
Tabela 10 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com alta concentração de pontos, utilizando a matriz **W**. 65
Tabela 11 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com diferentes variabilidades, utilizando o K-*nn* com a distância de Gowda-Diday para o cálculo da matriz $\mathbf{\Lambda}$. 71
Tabela 12 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com diferentes variabilidades, utilizando o k-*nn* com a distância de City block para o cálculo da matriz $\mathbf{\Lambda}$. 72
Tabela 13 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com diferentes variabilidades, utilizando o k-*nn* com a distância euclideana para o cálculo da matriz $\mathbf{\Lambda}$. 73
Tabela 14 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com diferentes variabilidades, utilizando o k-*nn* com a distância de Hausdorff para o cálculo da matriz $\mathbf{\Lambda}$. 74
Tabela 15 – Conjunto de dados simbólicos intervalar para o preço de aluguéis na cidade de Munique. 76
Tabela 16 – Valores para o I, $IASi$, $IASi^*$ e $IASi^*$(amplitude=0), com a matriz **W**. . 77

Tabela 17 – Valores obtidos para os índices I, $IASi$, $IASi^*$ e $IASi^*$(amplitude=0), quando a cidade é dividida em duas regiões (Grupos), utilizando a matriz **W**. 78
Tabela 18 – Valores para o I, $IASi$, $IASi^*$ e $IASi^*$(amplitude$= 0$), quando a cidade é dividida em quatro grupos, utilizando a matriz de pesos **W**. 79
Tabela 19 – Valores para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, utilizando uma distância intervalar para o cálculo da matriz $\mathbf{\Lambda}$. 79
Tabela 20 – Valores do $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$ para dados da COVID-19, utilizando a matriz **W**. 83
Tabela 21 – Índices de autocorrelação espacial para dados de COVID-19, utilizando a matriz de pesos $\mathbf{\Lambda}$. 83
Tabela 22 – Conjunto de dados de COVID-19. 87
Tabela 23 – Desempenho do modelo de regressão linear intervalar (MMREE) e o desvio padrão para diferentes configurações de dados intervalares. 93
Tabela 24 – Desempenho do modelo CAR intervalar (MMREE) e o desvio padrão para diferentes configurações de dados intervalares. 94
Tabela 25 – Performance para o modelo de regressão linear CAR intervalar (MMREE) para diferentes configurações de dados, no qual a região de vizinhança é identificada usando uma distância clássica. 95
Tabela 26 – Performance para o modelo de regressão linear CAR intervalar (MMREE) para diferentes configurações de dados, no qual a região de vizinhança é identificada usando uma distância intervalar. 96
Tabela 27 – Índices de autocorrelação espacial para dados clássicos. 99
Tabela 28 – Conjunto de dados simbólico intervalar para o estado de Pernambuco. . . . 100
Tabela 29 – Avaliação de desempenho dos modelos para o estado de Pernambuco, com base na MMREE. 102
Tabela 30 – Índices de autocorrelação espacial para a região Nordeste do Brasil. 103
Tabela 31 – Conjunto de dados simbólicos intervalares para a região Nordeste do Brasil. 104
Tabela 32 – Avaliação do desempenho dos modelos para a região do Nordeste brasileiro. 105

LISTA DE ABREVIATURAS E SIGLAS

CAR	Autoregressivo Condicionalmente
I	Índice de Moran
IASi	Índice de Autocorrelação Simbólico intervalar com base na representação mínimo-máximo
IASi*	Índice de Autocorrelação Simbólico intervalar com base na representação centro-amplitude
kcn	K-centróides mais próximos
knn	K-vizinhos mais próximos
MMREE	média da magnitude relativa do erro da estimativa
SAR	Autoregressivo Simultâneo
SDA	Análise de Dados Simbólicos

SUMÁRIO

1 INTRODUÇÃO 16
1.1 MOTIVAÇÃO 18
1.2 OBJETIVOS 19
1.2.1 Questões de Pesquisa 20
1.3 PRINCIPAIS CONTRIBUIÇÕES 21
1.4 ORGANIZAÇÃO DA TESE 21
2 FUNDAMENTAÇÃO TEÓRICA 23
2.1 ANÁLISE DE DADOS SIMBÓLICOS 23
2.1.1 Tipos de Dados 26
2.1.2 Estatísticas Descritivas 29
2.1.3 Distâncias Intervalares 30
2.1.4 Análise de Regressão intervalar 34
2.1.4.1 Método do Mínimo e Máximo 36
2.1.4.2 Método do Centro e Amplitude 37
2.2 ANÁLISE DE DADOS ESPACIAIS 40
2.2.1 Índice de Autocorrelação Espacial 45
2.2.2 Modelos de Regressão para Dados Espaciais 48
3 ÍNDICES DE AUTOCORRELAÇÃO ESPACIAL INTERVALAR . . . 49
3.1 ÍNDICES INTERVALARES 50
3.1.1 Cálculo do peso binário 52
3.1.2 Validação Bootstrap 52
3.2 AVALIAÇÃO EXPERIMENTAL 53
3.2.1 Resultados para os dados espaciais não-autocorrelacionados 57
3.2.2 Resultados para os dados espaciais autocorrelacionados 60
3.3 DADOS INTERVALARES REAIS 75
3.3.1 Aplicação I 75
3.3.2 Aplicação II 79
3.4 CONSIDERAÇÕES FINAIS 83
4 REGRESSÃO LINEAR PARA DADOS ESPACIAIS NO CONTEXTO INTERVALAR 85

4.1 DADOS ESPACIAIS DE INTERVALO SIMBÓLICO 86
4.2 MODELO DE REGRESSÃO ESPACIAL PARA DADOS DE VALOR SIMBÓLICO INTERVALAR 87
4.2.1 Representação Mínimo e Máximo (Min-Max) **88**
4.2.2 Representação Centro e Amplitude (Cen; Amp) **88**
4.3 AVALIAÇÃO EXPERIMENTAL COM DADOS SINTÉTICOS INTERVALARES 89
4.4 APLICAÇÕES COM CONJUNTOS DE DADOS AGREGADOS 98
4.4.1 Aplicação I: Dados de mortalidade infantil **98**
4.4.2 Aplicação II: Renda Per Capita **102**
4.5 CONSIDERAÇÕES FINAIS 106
5 CONCLUSÕES 107
REFERÊNCIAS 109
APÊNDICE A – ALGUMAS DEFINIÇÕES 115
APÊNDICE B – PUBLICAÇÕES 116

1 INTRODUÇÃO

A dinâmica dos fenômenos ocorridos no espaço geográfico constitui um dos grandes desafios para a elucidação de questões centrais em diversas áreas do conhecimento, tais como ecologia, epidemiologia e ciências biológicas. A análise de dados espaciais está interessada em descrever os padrões existentes nos dados espaciais e estabelecer, preferencialmente de forma quantitativa, relações entre as diferentes variáveis geográficas. A análise de processos pontuais (RIPLEY, 1984; GETIS; BOOTS, 1978; ROGERS, 1974) é um bom exemplo de um conjunto de técnicas que foram usadas em vários fenômenos geográficos nas últimas três décadas. A análise de dados espaciais fornece métodos para a criação e interpretação de mapas, bem como é importante para auxiliar na tomada de decisões (BIVAND et al., 2008).

Uma das ideias mais importantes na análise de dados espaciais é o conceito de autocorrelação espacial. Esta quantidade mede a similaridade (ou dissimilaridade) entre objetos ou atividades em uma determinada área geográfica, com base em sua proximidade entre as áreas vizinhas (FISCHER; WANG, 2011). Sua existência pode ser vista a partir da proposição de Tobler (1970) chamada de 'a primeira lei da geografia: tudo está relacionado a todo o resto, mas coisas próximas estão mais relacionadas que coisas mais distantes'. Neste contexto, para um dado evento observado, existe uma localização espacial correspondente que pode ser descrita pela sua posição no mapa ou por sistemas de referência de coordenadas geográficas (BIVAND et al., 2008). A presença de autocorrelação espacial pode ser identificada por várias técnicas que lidam simultaneamente com informações de localização e atributos.

O objetivo do cálculo da autocorrelação espacial é verificar a existência ou não de um padrão espacial em uma determinada região. Dependendo da existência ou não de um padrão espacial, é possível descrever o comportamento de uma variável aleatória em relação ao espaço geográfico onde ocorrem os fenômenos de interesse. Medidas de autocorrelação espacial podem ser úteis para indicar o grau de similaridade entre diferentes regiões. Por exemplo, consideremos a seguinte situação: imagine que estejamos estudando índices relativos à produção agrícola em regiões vizinhas que historicamente apresentam comportamentos semelhantes. Então, em determinado período, repetimos a análise e encontramos uma mudança abrupta nesse comportamento, indicando agora a ausência de autocorrelação espacial na região em estudo. Pode ser resultado de algum problema na produção de uma das regiões, devido ao clima, ou mesmo a presença de alguma praga nas fazendas. Dessa forma, a autocorrelação

espacial poderia nos auxiliar na tomada de decisões para solucionar esses problemas e fornecer um quantificador para comparar o desempenho da produção entre as regiões analisadas.

Estudos envolvendo o conceito de autocorrelação espacial têm sido relatados em diversas áreas. Recentemente Cima et al. (2018), ao realizar um estudo sobre a produção de soja no estado do Paraná, Brasil, identificou regiões com baixa e alta produção total de grãos. Outra contribuição recente analisou a distribuição dos casos de COVID-19 e leitos de terapia intensiva no Estado do Ceará, Brasil (PEDROSA; ALBUQUERQUE, 2020).

Atualmente, a sociedade precisa lidar com uma imensa quantidade de dados nos mais diversos ramos da ciência. Em particular, em muitas situações, é necessário o tratamento de dados que carregam informação sobre o espaço geográfico. Tais dados são ditos georreferenciados. O estudo de dados espaciais é uma linha de pesquisa bastante ativa e promissora, com novos métodos e técnicas sendo desenvolvidos nos dias atuais. Dados em que temos informação sobre as suas coordenadas geográficas são também chamados de dados georreferenciadas. Nesse tipo de dado o aspecto espacial do fenômeno é sua característica mais importante. Qualquer nova proposta teórica precisa preservá-la, ou seja, precisa levar em consideração a localização geográfica em que a informação está inserida. Nesse contexto, um outro aspecto importante a ser considerado diz respeito à descrição da variabilidade existente em uma determinada área. Em outras palavras, é necessário descrever propriamente como uma dada variável se comporta com respeito à sua característica geográfica. Por exemplo, é pertinente estudar como uma dada variável de interesse varia quando as coordenadas geográficas variam, dentro de uma dada região. Com essa finalidade, buscamos explorar novas metodologias que levem em consideração a localização do objeto observado e que possam ser utilizadas em contextos mais gerais, envolvendo variáveis intervalares, por exemplo. Aplicações importantes de tal realização podem incluir estudos epidemiológicos, sociais e econômicos. Neste estudo, estamos focados em propor uma nova estrutura a respeito do tratamento de grandes bases de dados espaciais. Mais especificamente, vamos empregar a estrutura da Análise de Dados Simbólicos (do inglês, Symbolic Data Analysis, abreviado como SDA).

A Análise de Dados Simbólicos (SDA) busca reduzir a quantidade de informações, nas quais os conjuntos de dados são agregados em tamanhos de dados menores e mais gerenciáveis, de modo que seja possível acomodar informações sobre a variabilidade existente em um conjunto de dados de interesse. Após a agregração dos dados, as classes são consideradas novas unidades estatísticas com maior nível de generalização do que os indivíduos (DIDAY, 2016). Além disso, a SDA fornece uma maneira de pensar os dados com o estabelecimento de

classes de indivíduos/objetos descritos por dados que permitem levar em consideração qualquer variabilidade inerente à um conjunto de dados. Tal variabilidade pode ser expressa por dados simbólicos tais como intervalos, lista de valores/categorias, histogramas ou distribuições de probabilidades. Segundo Bertrand e Goupil (2000), a SDA abre um vasto domínio de pesquisa e aplicações, estendendo métodos da análise de dados padrão para dados simbólicos. Esse paradigma fornece uma nova forma de pensar em ciência de dados, na qual a variável analisada é representada por um conjunto de classes de indivíduos.

Então, uma questão pertinente consiste em estender o conceito de autocorrelação espacial aos dados simbólicos. Tal extensão é baseada na possibilidade de agregar dados espaciais de cada região em grupos de interesse, resultando em um banco de dados mais condensado e considerando a variabilidade existente em cada conjunto de regiões. No contexto da análise de dados clássicos, diferentes indicadores podem ser usados para mensurar a autocorrelação espacial, todos baseados na mesma ideia: verificar como a dependência espacial muda, comparando valores entre uma amostra e seus vizinhos. Dentre as metodologias encontradas, o índice de Moran é uma das mais utilizadas na literatura para medir a autocorrelação espacial. Alternativas ao índice clássico de Moran foram propostas em Chen (2013), no entanto, a variabilidade existente não é abordada. Uma nova proposta consiste em considerar como a variabilidade existente na região em estudo impactará no valor do índice.

Em dados espaciais, quando está presente a autocorrelação espacial, as estimativas do modelo de regressão devem incorporar essa estrutura espacial, uma vez que a dependência entre as observações altera o poder explicativo do modelo (MONTEIRO et al., 2004). Dessa maneira, buscamos utilizar modelos de regressão que incorporam os efeitos espaciais de forma global, ou seja, como um único parâmetro. Modelos de regressão que levam em consideração o grau de conectividade existente entre as regiões de vizinhança podem ser encontrados na literatura em Bivand et al. (2008), Banerjee, Carlin e Gelfand (2014), Schabenberger e Gotway (2017), Hoef, Hanks e Hooten (2018)

1.1 MOTIVAÇÃO

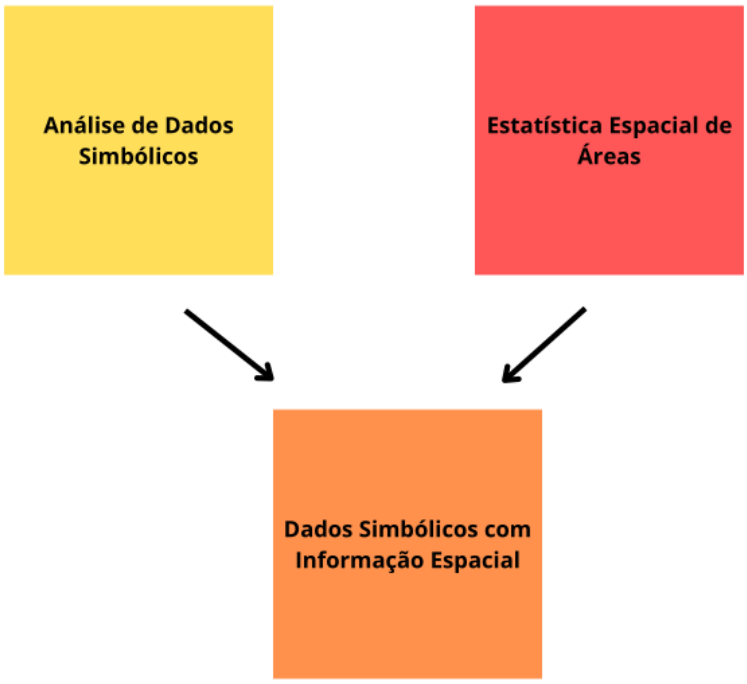
A ausência de métodos para lidar com dados simbólicos no contexto de dados georreferenciados é evidente na literatura de SDA, demandando atenção ao desenvolvimento de metodologias inovadoras. Um dos objetivos do presente estudo é apresentar um índice de autocorrelação para dados intervalares no contexto da análise de dados espaciais. Com essa

finalidade, generalizamos um índice de autocorrelação para variáveis da análise espacial, para o contexto de dados intervalares, no qual investigamos a autocorrelação espacial levando em consideração a variabilidade existente dentro de cada região. Uma outra proposta é ajustar um modelo de regressão espacial que leve em consideração a variabilidade existente na região no contexto de dados simbólicos intervalares. Mais precisamente, iremos adotar o modelo espacial autoregressivo condicionalmente no contexto de dados simbólicos do tipo intervalo.

1.2 OBJETIVOS

O objetivo geral deste trabalho é analisar dados simbólicos do tipo intervalo, no contexto de Estatística Espacial de Áreas. Dessa maneira, generalizamos um índice de autocorrelação espacial para dados intervalares e introduzimos um modelo de regressão espacial para dados intervalares. Desse modo, introduzimos uma nova contribuição à literatura de ambas as áreas, propondo uma metodologia que permite estudar conceitos e técnicas de ambas, agora de forma unificada. A Figura 1 esboça essa proposta.

Figura 1 – Objeto de pesquisa da tese.



Fonte: Elaborada pela autora.

1.2.1 Questões de Pesquisa

Um conjunto de perguntas de pesquisa que nortearam as suposições presentes neste trabalho foram:

1. Qual o impacto na autocorrelação espacial quando o índice analisado é intervalar? Isto é, como o índice se altera quando a variável de estudo é do tipo intervalo?

2. Um índice de autocorrelação espacial intervalar é capaz de captar a variabilidade existente na região, para o caso de uma variável intervalar?

3. Qual a interpretação do índice intervalar quando a matriz de pesos é calculada utilizando distâncias intervalares?

4. Qual o comportamento do índice de autocorrelação espacial intervalar em comparação com o seu análogo clássico?

5. Qual o desempenho do modelo de regressão linear para variáveis do tipo intervalo, quando os dados são autocorrelacionados? Ele consegue captar a dependência espacial?

Ao longo desse trabalho são descritos métodos, estudos de simulações e aplicações em conjundo de dados reais que buscam responder a essas perguntas.

1.3 PRINCIPAIS CONTRIBUIÇÕES

Ao longo do texto são apresentados estudos de simulações com diferentes valores para a correlação espacial, tanto no cenário de ponto, quanto intervalar. Diferentes formas para o cálculo da matriz de vizinhança são utilizadas buscando identificar a conectividade existente entre as regiões de forma representativa. Para a identificação dessa região de vizinhança, distâncias clássicas e intervalares foram consideradas. As principais contribuições deste trabalho são:

- Índice de autocorrelação espacial intervalar para diferentes cenários de variabilidade, utilizando a parametrização mínimo-máximo;
- Índice de autocorrelação espacial intervalar para diferentes cenários de variabilidade, utilizando a parametrização centro-amplitude;
- Aplicação do índice de autocorrelação espacial intervalar para dados de aluguéis;
- Aplicação do índice de autocorrelação espacial intervalar para dados de COVID-19;
- O desempenho do modelo de regressão linear intervalar para dados autocorrelacionados, sob diferentes parametrizações;
- O desempenho do modelo de regressão espacial intervalar para dados autocorrelacionados, sob diferentes parametrizações;
- Aplicações em conjuntos de dados reais utilizando o modelo espacial intervalar.

1.4 ORGANIZAÇÃO DA TESE

O Capítulo 2 apresenta uma breve revisão da SDA e os campos de análise de dados espaciais também são discutidos, no qual alguns conceitos fundamentais da análise de dados espaciais

são abordados. No Capítulo 3, propomos um índice de autocorrelação espacial para dados de intervalo com diferentes parametrizações. Nesse capítulo é realizada uma avaliação dos índices de autocorrelação aplicados a dados sintéticos utilizando valores intervalres. São apresentadas duas aplicações com dados reais: a primeira é referente ao preço do aluguel na cidade de Munique e a segunda refere-se ao número de casos de COVID-19 no Nordeste brasileiro. No Capítulo 4, é apresentado um modelo de regressão linear espacial para dados de intervalo. A ideia é analisar o desempenho desse modelo com modelos intervalares existentes na literatura. Estudos de simulação e aplicações em conjuntos de dados reias são considerados para avaliar o desempenho dessas abordagens. Finalmente, o Capítulo 5 apresenta as observações finais e as perspectivas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Muitos problemas associados com autocorrelação espacial e/ou heterogeneidade espacial não levam em consideração a variabilidade existente dentro de uma área estudada. Isso quer dizer que, em muitas situações, a amplitude de variação de um dado fenômeno não é considerada. Uma maneira de captar a variabilidade existente nessa área, consiste em estender métodos usuais de autocorrelação espacial para o caso envolvendo dados simbólicos, uma vez que esses dados nos permitem levar em consideração a heterogeneidade dentro de uma área geográfica. Além disso, o emprego de uma variável intervalar já permite, por definição, acomodar os valores máximo e mínimo para uma observação. Essa abordagem é um ponto em aberto na literatura de SDA. A seguir, apresentamos uma revisão dos principais conceitos envolvendo os campos de Análise de Dados Simbólicos e Análise de Dados Espaciais. Na análise de dados simbólicos são apresentados suas principais características, aplicações e uma revisão dos trabalhos mais relevantes desenvolvidos nesta área. Em seguida, apresenta-se conceitos e metodologias utilizadas na análise de dados georrefenciados.

2.1 ANÁLISE DE DADOS SIMBÓLICOS

Com a crescente demanda por informações, grandes conjuntos de dados precisam ser resumidos, nos quais o conjunto de dados resultante preserva o conhecimento dos dados originais de forma representativa (DIDAY, 2016). Uma consequência dessa redução é que o conjunto de dados pode ser representado de várias maneiras. Dentre as alternativas encontradas na literatura, a agregação de dados pode ser feita com a abordagem de dados simbólicos. A análise de dados simbólicos propõe uma abordagem alternativa para lidar com conjuntos de dados grandes e complexos, além de permitir a compressão desses conjuntos de dados em outros menores e mais gerenciáveis, retendo o conhecimento chave. Em um conjunto de dados simbólicos, os itens são descritos por variáveis simbólicas, como listas de valores (categóricos ou quantitativos), intervalos e distribuições (NOIRHOMME-FRAITURE, 2002).

A SDA é uma nova maneira de pensar sobre ciência de dados, estendendo a análise de dados padrão aos dados simbólicos. Essa abordagem busca aprimorar o estudo de uma população de indivíduos descritos por variáveis clássicas. Para descrever as classes, os valores simbólicos das variáveis são construídos a partir de observações fornecidas sobre a população base. Essas

classes podem ser descritas como variáveis simbólicas e possuem diferentes variabilidades (BILLARD; DIDAY, 2006a). Para obter as classes, é realizado um procedimento conhecido como agregação. A agregação das realizações de uma variável aleatória em um intervalo implica na perda da distribuição de probabilidade dos dados (PIMENTEL; SOUZA, 2014). Entretanto, resulta em um conjunto de valores, nos quais a variabilidade observada dentro de cada classe é extremamente importante. Agregar classes e descrevê-las como variáveis simbólicas resulta em várias vantagens: i) redução do número de indivíduos, ii) redução de variáveis, iii) redução da perda de informação, quando comparado aos dados clássicos (DIDAY, 2016).

Variáveis simbólicas têm variabilidade inerente. Por exemplo, uma variável com um valor de intervalo é um tipo de variável simbólica, uma vez que um intervalo expressa uma variabilidade dentro da classe que descreve (BERTRAND; GOUPIL, 2000). Esta é a forma mais comum de considerar uma variável simbólica na literatura e comumente encontrada na prática (KAO et al., 2014).

Diversas metodologias para dados de intervalo foram propostas na literatura de SDA (BILLARD; DIDAY, 2006c). Um estudo de revisão recente sobre este assunto pode ser acessado em Diday (2016). Estudos incluindo estatísticas descritivas podem ser encontrados em Bertrand e Goupil (2000), Billard e Diday (2006c). A técnica de análise de componentes principais no contexto de dados simbólicos é encontrada em Douzal-Chouakria, Billard e Diday (2011). Trabalhos relacionados a análise de agrupamentos ou análise discriminante podem ser estudados em Lauro, Verde e Palumbo (2000), Souza et al. (2004), Pimentel e Souza (2014), Li et al. (2016), Bao et al. (2018). Metodologias envolvendo modelos de regressão no contexto de dados simbólicos podem ser vistas em diversos trabalhos na literatura, dentre eles: Billard e Diday (2000), Neto e Carvalho (2008), Xu (2010), Domingues, Souza e Cysneiros (2010), Fagundes, Souza e Cysneiros (2013), Suresh, Brito e Dias (2020). A SDA pode ser encontrada em disversos campos da literatura, dentre elas, redes neurais (YANG; HUNG; CHEN, 2012); machine learning (DIDAY, 2018) e reconhecimento facial (OCHS; DIDAY; AFONSO, 2015).

Para uma compreensão mais ampla de uma variável simbólica, suponha que desejamos descrever as características referentes à produção de uma fruta (por exemplo, maçã) em uma determinada região. Podemos dizer 'O peso está entre 25 e 60 gramas e sua cor é vermelha ou verde, entretanto quando a cor é verde, o peso é inferior a 35 gramas'. No contexto de dados clássicos esse tipo de informação não é possível de ser organizada em uma base de dados. Uma forma viável de representar esses dados consiste em utilizar o conceito de dados simbólicos

(bidimensional), sendo representado como

$$[\text{Peso} = [25, 60]],\ [\text{Cor} = \text{vermelha, verde}] \text{ e } [\text{Se Cor} = \text{ verde, então Peso} < 35].$$

Segundo Billard e Diday (2006c), um 'ponto' de dados simbólicos assume valores como um hipercubo (ou hiper-retângulo, amplamente definido) ou um produto cartesiano de distribuições no espaço p-dimensional, sendo representados por listas de valores, histogramas, intervalos e assim por diante. Além disso, os dados simbólicos podem aparecer em diferentes situações no mundo real. Podem surgir de agregação de conjuntos de dados muito grandes ou existir de maneira natural.

Os dados simbólicos de natureza intervalar buscam resumir grandes conjuntos de dados relacionados à observações individuais para valores intervalares. Tipos de dados dessa natureza fornecem uma gama de valores variáveis e podem ser facilmente encontrados. Como exemplo, podemos citar a pressão sanguínea e os valores de temperatura. Outra possibilidade refere-se a dados contendo imprecisões advindas de medidas repetidas ou estimativa de intervalo de confiança, resultantes de algum experimento. Além disso, os dados de intervalos podem ser obtidos a partir de dados clássicos. Nesse caso, os dados de intervalo são criados agregando observações individuais clássicas das regiões e descrevendo cada uma por valores mínimo e máximo.

Em ciência de dados, frequentemente é necessário lidarmos com grandes conjuntos de dados. Uma forma para analisar esses dados consiste em agrupá-los os dados individuais em classes. No contexto de dados simbólicos, essas classes são consideradas como novas unidades de maior nível de generalização do que os indivíduos.

As variáveis simbólicas são variáveis que associam a cada classe um valor simbólico. De acordo com Diday (2016), uma 'tabela de dados simbólicos' pode ser representada por uma tabela no qual as classes de indivíduos são descritas por pelo menos uma variável. Para a compreensão de uma variável simbólica, considere o seguinte exemplo de dados simbólicos que é ilustrado pela Tabela 1, no qual, os indivíduos são os jogadores dos times da copa francesa e as classes de jogadores são equipes chamadas Paris, Lyon, Marselha e Bordeaux. De acordo com essa tabela, podemos encontrar diversos tipos de variáveis simbólicas. Como exemplo, a variável 'Peso', cujo valor é o intervalo de [mínimo; máximo] representando o peso dos jogadores. Uma outra variável do tipo "lista", que expressa a nacionalidade do jogadores. Uma outra representação pode ser obtida com a varável simbólica 'idade', que pode ser representada por um 'gráfico de barras variável'. E por fim, a variável que expressa a frequência

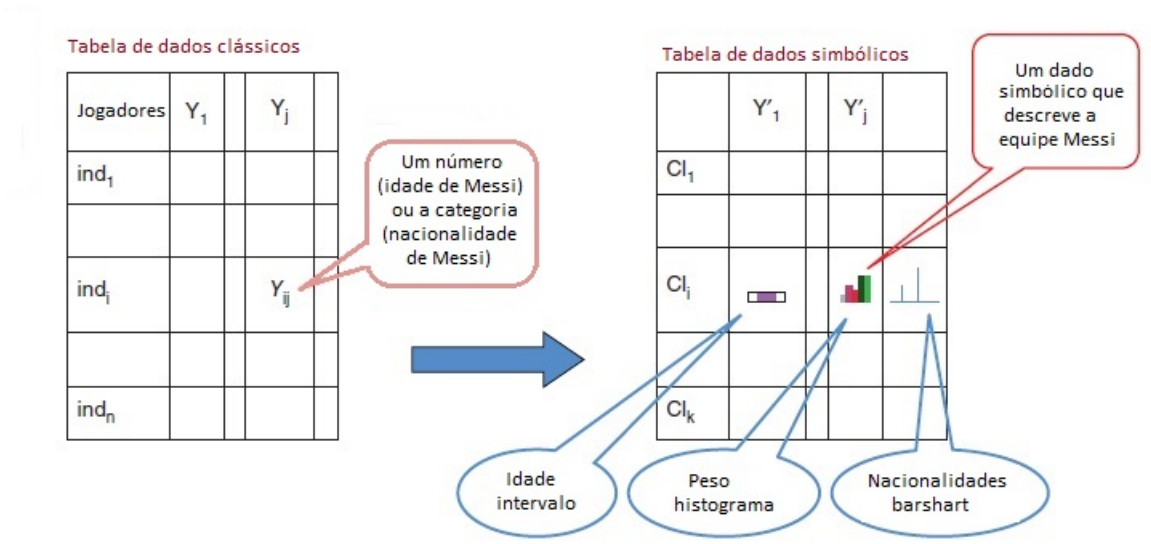
dos jogadores franceses nesta equipe entre todos os jogadores franceses de todas as equipes, sendo representado por gráficos de barras horizontais. Para nível de representação, considere as informações apresentadas na Tabela 1 sendo explanadas na Figura 2.

Tabela 1 – Exemplo de uma tabela de dados simbólicos com os times da copa francesa.

Times da copa francesa	**Peso**	**País de nacionalidade**	**Idade**	**Freq. de francenses entre times franceses**
Paris	[73; 85]	{França, Argentina, Senegal}	{(0) 30%, (1) 70%}	30
Lyon	[68; 90]	{França, Brasil, Itália }	{(0) 20%, (1) 65%, (2) 5%}	25
Marseille	[77; 85]	{França, Brasil, Algéria}	{(1) 40%, (2) 52%, (3) 8% }	28
Bordeaux	[80; 90]	{França, Argentina}	{(0) 40%, (1) 60% }	17

Fonte: Tabela adaptada de Diday (2016).

Figura 2 – Explanação de uma tabela de dados em que (X, Y) representam variáveis clássicas, para uma tabela de dados simbólicos (X', Y') em que X' e Y' descrevem um conjunto de variáveis simbólicas.



Fonte: Figura adaptada de Diday (2016).

2.1.1 Tipos de Dados

Ferramentas para analisar e extrair novos conhecimentos são desenvolvidas com base no tipo de dado a ser analisado. Essa análise pode ser realizada desenvolvendo metodologias específicas ou ampliando as ferramentas conhecidas. No contexto de dados simbólicos, diversas abordagens podem ser encontradas de acordo com os tipos de dados, como por exemplo: Estatística Descritiva, Análise de Componentes Principais, Análise de Regressão, Árvores de Decisão, Agrupamentos, Medidas de Dissimilaridades, entre outras.

Uma característica inerente aos dados simbólicos, é que eles são definidos em classes. Uma outra característica é que seus valores expressam a variabilidade entre os indivíduos dentro dessas classes. Segundo Billard e Diday (2006c), Diday e Noirhomme-Fraiture (2008), as variáveis simbólicas podem assumir uma distribuição, um intervalo, um histograma, uma lista de categorias ou números, etc.

As variáveis simbólicas podem ser obtidas a partir de variáveis clássicas e definidas como uma tabela de dados simbólicos. A ideia é que a SDA possa melhorar o estudo usual de qualquer população de indivíduos descrito por variáveis clássicas, adicionando um estudo complementar de classes de indivíduos descrito por variáveis simbólicas que expressam a variabilidade dessas classes (DIDAY, 2016).

Um tipo de variável comumente utilizada em SDA é a variável com valor de intervalo. Esse tipo de variável simbólica expressa por si só uma variabilidade dentro da classe que a descreve. Por exemplo, o [Mínimo, Máximo] ou o intervalo inter-quartil da variável aleatória 'idade', dentro de uma equipe considerada como uma classe de jogadores, (Tabela 1). Para representação dessa variável, considere a definição a seguir.

Definição 1 *Seja* $E = \{R_1, \ldots, R_m\}$ *um conjunto de* m *classes descritas pela variável de intervalo* Z*. Cada item de* E *é representado pelo intervalo* $z_i = [a_i, b_i] \in \mathbb{I}[a, b] : a, b \in \mathbb{R}$ *e* $a \leq b$*. Um intervalo é definido como um conjunto de números reais apresentados por*

$$(a, b),\ [a, b],\ [a, b)\ \textit{ou}\ (a, b] \tag{2.1}$$

As variáveis com valor de intervalo podem ocorrer naturalmente quando, por exemplo, não é possível obter medições precisas em indivíduos. Os intervalos podem ser construídos a partir de dados pontuais por meio da agregação de dados clássicos.

Tabela 2 – Conjunto de dados clássicos com gastos (em dólares) do cartão de crédito.

i	Nome	Mês	Alimentação	Social	Viagem	Gasolina	Roupas
1	Jon	Fevereiro	23,65	14,56	218,02	16,79	45,61
2	Leigh	Maio	28,47	8,99	141,60	21,74	86,04
3	Leigh	Julho	30,86	9,55	193,14	24,26	95,68
4	Tom	Julho	24,13	15,97	190,40	35,71	20,02
5	Jon	Abril	23,40	11,61	179,38	23,73	48,89
6	Jon	Novembro	23,11	16,71	178,78	20,55	47,96
7	Leigh	Setembro	32,14	12,34	165,65	17,62	66,40
8	Leigh	Agosto	25,92	20,78	201,18	32,97	70,96
9	Leigh	Novembro	31,52	16,62	177,50	20,95	71,18
10	Jon	Novembro	23,11	14,41	179,86	20,53	51,49
11	Jon	Novembro	22,80	11,35	184,55	20,94	50,36
12	Leigh	Setembro	32,83	13,93	158,65	17,04	69,41
13	Leigh	Novembro	31,13	12,82	179,57	20,67	69,01
14	Tom	Agosto	23,01	13,20	220,52	29,44	18,09
15	Jon	Dezembro	21,09	9,90	180,66	22,95	47,87
16	Leigh	Agosto	30,90	13,29	202,22	32,29	68,71
17	Leigh	Dezembro	37,36	15,63	184,22	20,32	71,74
18	Tom	julho	24,25	15,71	149,01	30,68	21,75
19	Tom	Abril	21,83	14,95	154,43	30,48	21,09
20	Jon	Janeiro	25,94	12,38	197,90	20,06	47,09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Fonte: Tabela adaptada de Billard e Diday (2006c).

Para observar essa característica, considere a Tabela 2, no qual uma lista parcial das despesas com cartão de crédito por um grupo de indivíduos em um período de 12 meses é apresentada. O emissor do cartão está interessado nas despesas discriminadas de cada indivíduo (em dólares) para as variáveis: alimentação, social, viagem, gasolina e roupas, mensalmente por pessoa. Se o interesse do emissor está na categoria de pessoa-mês e os padrões de gastos relacionados, a Tabela 3 resume as informações da Tabela 2 de acordo com o critério de interesse. De acordo com a Tabela 3, o valor para cada observação é caracterizada por um intervalo. Por exemplo, no mês de janeiro, Os gastos de 'Jon' com alimentação assumem valores no intervalo [20,81; 29,38], os de Tom são [23,28; 30,00].

Intervalos gerados na forma $[a, a]$, também chamados de intervalos degenerados, são equivalentes a números reais. Segundo Billard e Diday (2006a), esse é um caso particular de uma observação com valor de intervalo.

Tabela 3 – Conjunto de dados intervalares com gastos do cartão de crédito.

Nome - Mês	Alimentação	Social	Viagem	⋯	Roupas
Jon – Jan	[20,81; 29,38]	[9,74; 18,86]	[192,33; 205,23]	⋯	[44,28; 53,82]
Jon – Fev	[21,44; 27,58]	[10,86; 18,01]	[214,98; 229,63]	⋯	[50,51; 63,57]
⋮	⋮	⋮	⋮	⋮	⋮
Tom – Jan	[23,28; 30,00]	[8,67; 18,31]	[193,53; 206,53]	⋯	[15,51; 25,66]
Tom – Fev	[20,61; 28,66]	[10,66; 17,20]	[195,53; 203,83]	⋯	[12,99; 24,88]
⋮	⋮	⋮	⋮	⋮	⋮
Leigh – Jan	[25,59; 35,33]	[7,07; 19,00]	[194,12; 207,05]	⋯	[61.47; 75.43]
Leigh – Fev	[31,30; 40,80]	[9.05; 24,44]	[212,76; 227,43]	⋯	[71,63; 85,58]
⋮	⋮	⋮	⋮	⋮	⋮

Fonte: Tabela adaptada de Billard e Diday (2006c).

Dados simbólicos de natureza intervalar também são encontrados na literatura. Um exemplo pode ser observado na Tabela 4, na qual dados 'naturalmente' simbólicos do tipo intervalo são apresentados. Esses dados referem-se as temperaturas mensais, mínimas e máximas, registradas em 60 estações meteorológicas na China para o ano de 1988 (BILLARD; DIDAY, 2006c).

Tabela 4 – Exemplo de uma tabela simbólica com dados 'naturalmente' intervalares.

Estações	Janeiro	Fevereiro	⋯	Dezembro
AnQing	[1,8; 7,1]	[2,1; 7,2]	⋯	[4,3; 11,8]
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
ZhoJing	[2,7; 8,4]	[2,7; 8,7]	⋯	[5,1; 13,3]

Fonte: Tabela adaptada de Billard e Diday (2006c).

Dessa maneira, podemos ver que cada célula de uma tabela de dados simbólicos pode ser representada por diferentes tipos de dados (Tabela 1), ou por variáveis simbólicas do mesmo tipo (Tabela 4).

2.1.2 Estatísticas Descritivas

As estatísticas descritivas básicas para uma variável aleatória incluem histogramas de frequência, médias e variâncias amostrais. No cenário de variáveis aleatórias com valor de intervalo, as estatísticas descritivas são obtidas analogamente àquelas para variáveis com valores múltiplos; ver Bertrand e Goupil (2000).

Metodologias para encontrar o histograma e as probabilidades de intervalo para variáveis com valor de intervalo podem ser encontradas em Carvalho (1995). Em Oliveira et al. (2018) é apresentada a média, variância e covariância para variáveis simbólicas do tipo intervalo, no qual utiliza-se aritmética intervalar para mostrar algumas propriedades. Para o cálculo dessas estatísticas descritivas, considere as definições apresentadas em Billard e Diday (2006c), em que

Definição 2 *Para uma variável aleatória Z com valor de intervalo, a* ***média amostral*** *simbólica é dada por*

$$\bar{Z} = \frac{1}{2m} \sum_{k \in E} (b_k + a_k), \tag{2.2}$$

onde b_k *e* a_k *são os limites superiores e inferiores, respectivamente. A* ***variância amostral*** *é expressa por*

$$S^2 = \frac{1}{3m} \sum_{k \in E} (b_k^2 + b_k a_k + a_k^2) - \frac{1}{4m^2} \left[\sum_{k \in E} (b_k + a_k) \right]^2. \tag{2.3}$$

As demonstrações para as Equações 2.2 e 2.3 podem ser encontradas na literatura em Billard e Diday (2006a).

2.1.3 Distâncias Intervalares

Os índices de autocorrelação espacial baseiam-se no grau de conectividade de uma região com a sua vizinhança. Assim, é necessário a adoção de critérios que permitam classificar se uma dada região deve ser considerada ou não como vizinha da região de interesse. Para isto, precisamos de uma definição de uma distância que quantifique a similaridade entre as regiões. Tal medida deve nos indicar a proximidade entre duas regiões.

Na literatura existem diversas medidas de similaridade ou dissimilaridade para dados simbólicos. Nesse contexto, iremos abordar distâncias para dados com valor de intervalo. Vamos denotar um intervalo u para a observação w_u por

$$\xi_u = ([a_{uj}, b_{uj}], j = 1, \cdots, p), u = 1, \cdots, m. \tag{2.4}$$

Nesse estudo consideramos as seguintes distâncias da literatura de SDA para o cálculo da matriz de pesos $\mathbf{\Lambda}$: a distância Gowda-Diday gowda1991symbolic; a distância de City block

(SOUZA; CARVALHO, 2004); a distância de Hausdorff (CHAVENT; LECHEVALLIER, 2002) e a distância Euclideana (BILLARD; DIDAY, 2006c). Para o cálculo dessas distâncias, considere as definições à seguir.

Definição 3 *A medida de **distância de Gowda-Diday** entre duas observações intervalares w_{u_1} e w_{u_2} da forma da Equação 2.4 são dadas por*

$$D(w_{u_1}, w_{u_2}) = \sum_{j=1}^{p} D_j(w_{u_1}, w_{u_2}), \tag{2.5}$$

no qual para a variável Y_j, a distância é expressa por

$$D_j(w_{u_1}, w_{u_2}) = \sum_{k=1}^{3} D_{jk}(w_{u_1}, w_{u_2}), \tag{2.6}$$

com

$$D_{j1}(w_{u_1}, w_{u_2}) = ||b_{u_1j} - a_{u_1j}| - |b_{u_2j} - a_{u_2j}||/k_j, \tag{2.7}$$

onde

$$k_j = |Max(b_{u_1j}, b_{u_2j}), Min(a_{u_1j}, a_{u_2j})|,$$

é o comprimento de toda a distância medida por w_{u1} e w_{u2}; com

$$D_{j2}(w_{u_1}, w_{u_2}) = (|b_{u_1j} - a_{u_1j}| + |b_{u_2j} - a_{u_2j}| - 2I_j)/k_j, \tag{2.8}$$

onde I_j é o comprimento da interseção dos intervalos $[a_{u_1j}, b_{u_1j}]$ e $[a_{u_2j}, b_{u_2j}]$, ou seja,

$$I_j = |Max(a_{u_1j}, a_{u_2j}) - Min(b_{u_1j}, b_{u_2j})|,$$

se os intervalos se sobrepõem, e $I_j = 0$ caso contrário; e com

$$D_{j3}(w_{u_1}, w_{u_2}) = |a_{u_1j} - a_{u_2j}|/|\mathcal{Y}_j|, \tag{2.9}$$

onde $\mathcal{Y}_j$ é o tamanho total em $\mathcal{Y}$ coberto pelos valores observados de Yj, ou seja,

$$|\mathcal{Y}_j| = \max_u(b_{uj}) - \min_u(a_{uj}). \tag{2.10}$$

Para ilustração do cálculo dessa distância, considere os dados da Tabela 5. Inicialmente, calculamos a distância de Gowda-Diday para a variável aleatória $Y_1 =$ altura. Utilizando a Equação 2.7, obtemos o componente de distância entre os cavalos machos (CavaloM) e fêmeas (CavaloF) como

$$D_{11}(\text{CavaloM, CavaloF}) = \frac{|(180-120)-(160-158)|}{|Max(180;160)-Min(120;158)|} = 0,967$$

Tabela 5 – Dados veterinários.

$\mathbf{w}_u$	**Animal**	**Altura**	**Peso**
w_1	CavaloM	[120,0; 180,0]	[222,2; 354,0]
w_2	CavaloF	[158,0; 160,0]	[322,0; 355,0]
w_3	UrsoM	[175,0; 185,0]	[117,2; 152,0]
w_4	VeadoM	[37,9; 62,9]	[22,2; 35,0]
w_5	VeadoF	[25,8; 39,6]	[15,0; 36,2]
w_6	CachorroF	[22,8; 58,6]	[15,0; 51,8]
w_7	CoelhoM	[22,0; 45,0]	[0,8; 11,0]
w_8	CoelhoF	[18,0; 53,0]	[0,4; 2,5]
w_9	Gato	[40,3; 55,8]	[2,1; 4,5]
w_{10}	Gata	[38,4; 72,4]	[2,5; 6,1]

Fonte: Tabela adaptada de Billard e Diday (2006c).

Da mesma forma, substituindo na Equação 2.8, obtemos

$$D_{12}(\text{CavaloM, CavaloF}) = \frac{|180-120|+|160-158|-2|160-158|}{|180-120|} = 0,967$$

e substituindo na Equação 2.9, encontramos

$$D_{13}(\text{CavaloM, CavaloF}) = \frac{|120-158|}{Max(180;160;185)-Min(120;158;175)} = 0,584$$

Portanto, a distância Gowda-Diday para a variável Y_1 é

$$D_1(\text{CavaloM, CavaloF}) = 0,967+0,967+0,584 = 2,518.$$

De acordo com a literatura, a **distância de City block** e a **distância Euclideana** são casos particulares da distância de Minkowski. Dessa maneira, podemos escrever formalmente as seguintes definições.

Definição 4 *A **distância de Minkowski generalizada** de ordem q para objetos com valor de intervalo w_{u_1} e w_{u_2} é*

$$d_q(w_{u_1}, w_{u_2}) = \{\sum_{j=1}^{p} w_j^*[\phi_j(w_{u_1}, w_{u_2})]^q\}^{1/q}, \tag{2.11}$$

onde w_j^* é uma função de peso associada a Y_j e

$$\phi_j(w_{u_1}, w_{u_2}) = |w_{u_1 j} \oplus w_{u_2 j}| - |w_{u_1 j} \otimes w_{u_2 j}| + \gamma(2|w_{u_1 j} \otimes w_{u_2 j}| - |w_{u_1 j}| - |w_{u_2 j}|),$$

em que, |A| é o tamanho do intervalo A =[a, b] e $0 \leq \gamma \leq 0,5$ é uma constante pré-especificada.

Dessa maneira, quando consideramos $q = 1$ na Equação 2.11 a ***distância de city block*** *é expressa por*

$$d_q(w_{u_1}, w_{u_2}) = \sum_{j=1}^{p} c_j w_j^* \phi_j(w_{u_1}, w_{u_2}), \tag{2.12}$$

no qual $c_j > 0$, $\sum c_j = 1$, onde a ***distância euclideana normalizada*** *é uma distância de Minkowski de ordem q = 2, ou seja,*

$$d(w_{u_1}, w_{u_2}) = \left(\frac{1}{p} \sum_{j=1}^{p} w_j^* [\phi_j(w_{u_1}, w_{u_2})]^2 \right)^{1/2}. \tag{2.13}$$

Para a ilustração do cálculo dessa distância, considere a Tabela 5 e as três primeiras classes. Assim,

$$\phi_1(\text{CavaloM, UrsoM}) = (185 - 120) - (180 - 175) + \gamma(2 \times 5 - 60 - 10) = 60 + \gamma(-60).$$

Considerando $\gamma = 0,5$ e substituindo na distância de Minkowski encontramos o valor 30. Para a distância de city block e euclidiana os valores foram $183,5$ e $156,4$, respectivamente.

Uma outra distância comumente utilizada na área de SDA é a distância de Hausdorff. Abordagens envolvendo essa métrica podem ser encontradas em: Carvalho et al. (2006), Carvalho e Simões (2017) e Li et al. (2016). A seguir, escrevemos sua definição.

Definição 5 *A* ***distância de Hausdorff*** *para uma variável Y_j entre dois valores observados ξ_{u_1} e ξ_{u_2} é expressa por*

$$\phi_j(w_{u_1}, w_{u_2}) = Max[|a_{u_1 j} - a_{u_2 j}|, |b_{u_1 j} - b_{u_2 j}|]. \tag{2.14}$$

Como um exemplo do cálculo dessa distância considere os dados da Tabela 5. A distância de Hausdorff para as observações de cavalos e ursos machos considerando a variável Y_1 é dada por

$$\phi_1(CavaloM, UrsoM) = Max[|120 - 175|, |180 - 185|] = 55.$$

No contexto de análise de dados simbólicos há uma série de medidas de distância para dados com valor de intervalo, algumas das quais possuem análogas àquelas para observações com vários valores, enquanto outras são exclusivas para dados com valor de intervalo.

2.1.4 Análise de Regressão intervalar

Sabe-se que a Análise de Dados Simbólicos é uma área de pesquisa e aplicação relacionada ao aprendizado de máquina e estatística, que fornece ferramentas para descrever geo-objetos, permitindo lidar com diferentes tipos de variáveis que levam em conta a variabilidade existente nos dados. Aqui, no contexto dos modelos de regressão, nos interessa modelar a relação existente entre um conjunto de variáveis, tendo em conta a sua natureza simbólica e incluindo informação sobre a posição geográfica.

Na estrutura de modelos de regressão para dados de intervalo simbólico, vários modelos foram introduzidos em Billard e Diday (2006b). A maioria desses modelos considera que seus parâmetros são estimados pelo método dos mínimos quadrados. Em Billard e Diday (2000), por exemplo, funções de covariância e correlação são introduzidas para dados simbólicos com valores de intervalo, e seus termos associados são então usados para ajustar um modelo de regressão linear.

Na literatura, também podemos encontrar a análise de regressão para dados simbólicos do tipo intervalar, sendo as medidas de centro e amplitude as mais apropriadas para representar dados intervalares (NETO; CARVALHO, 2008). Esse modelo é uma extensão da regressão linear múltipla padrão com base na estimativa de mínimos quadrados para dados de intervalo. Nesse contexto, Domingues, Souza e Cysneiros (2010) introduziram o primeiro modelo linear robusto para dados intervalares, assumindo uma distribuição de probabilidade simétrica para erros de centro e amplitude dos intervalos e permitindo lidar com intervalos discrepantes.

Em Neto, Cordeiro e Carvalho (2011), foram desenvolvidos modelos de regressão linear simbólica bivariada para representações de centro e amplitude e intervalos mínimos e máximos baseados no Modelo Linear Generalizado (MLG). Um modelo de regressão robusto para dados de intervalo com base na representação de centro e intervalo usando mínimos quadrados reponderados iterativamente e definições de outlier de intervalo com base no centro e/ou intervalo foi proposto em Fagundes, Souza e Cysneiros (2013). Modelos de regressão restritos ao intervalo foram apresentados com base em mínimos quadrados ordinários (NETO; CARVALHO, 2010). Fagundes, Souza e Soares (2016) introduziram a regressão quantílica intervalar (IQR) para dados simbólicos do tipo intervalo, representados pelos dados do centro e amplitude.

Um modelo de regressão linear baseado na representação mínimo-máximo, que utiliza uma transformação Box-Cox intervalar para intervalo de resposta variável quando se verifica que a coerência matemática não é garantida, é introduzido em Souza et al. (2017). Hao e Guo

(2017) propôs um modelo de centro e amplitude para regressão de dados simbólicos com valores intervalares para resolver o problema de coerência matemática referente à estimação intervalar. Em Silva, Souza e Cysneiros (2019), foi proposto um modelo de regressão linear para variáveis poligonais simbólicas que também considera a representação do centro e amplitude para os dados de entrada. Esse modelo é chamado de regressão linear poligonal (PLR), e os parâmetros desconhecidos são estimados pelo método dos mínimos quadrados.

No contexto de modelos não lineares, Lim (2016) apresentou um modelo aditivo não paramétrico para análise de dados de valores de intervalos com base em dados de centro e intervalo. O modelo aditivo é definido por uma função paramétrica e um termo aditivo baseado em funções suaves desconhecidas. Neto e Carvalho (2017) propuseram um modelo de regressão não linear para dados de valores intervalares que considera dois modelos de regressão não lineares independentes para o centro e amplitude dos intervalos, respectivamente. Outra proposta é dada em Reyes, Souza e Cysneiros (2017), no qual é apresentado um único modelo não linear para dados intervalares baseado na família simétrica de distribuições, em que a estimação de parâmetros e a previsão de novos padrões são menos sensíveis na presença de outliers intervalares. Em Carvalho, Neto e Ferreira (2017) uma regressão linear robusta é proposta usando um kernel exponencial para penalizar outliers no centro e amplitude, e as estimativas são baseadas em mínimos quadrados.

Modelos não paramétricos no contexto de abordagens de regressão não paramétrica baseadas em kernel podem ser encontrados na literatura SDA. Fagundes, Souza e Cysneiros (2014) propuseram duas famílias de modelos de regressão baseados em kernel para dados de intervalo usando representação central. Uma vantagem deste método sobre os modelos paramétricos para dados de intervalo é que o uso da função kernel para estimar dados de intervalo garante coerência matemática. Um artigo recente apresenta uma nova abordagem de três estágios para previsão de séries temporais com base em dados quartis que combina agregação, seleção de protótipos e previsão na estrutura de Análise de Dados Simbólicos (REYES; SOUZA; OLIVEIRA, 2021).

Esta seção apresentará alguns métodos de regressão para dados simbólicos do tipo intervalo baseados no método dos mínimos quadrados sem fazer suposição de distribuição de probabilidade para os erros.

2.1.4.1 *Método do Mínimo e Máximo*

O método do mínimo e máximo proposto por Billard e Diday (2002) é caracterizado por dois modelos diferentes: em um deles, o limite inferior da resposta se relaciona aos limites inferiores dos regressores, enquanto no outro, o limite superior da resposta depende dos limites superiores dos regressores. A modelagem para os limites inferior e superior da resposta são dadas por:

$$\dot{y}_i = \beta_0^{low} + \sum_{j=1}^{p} \beta_j^{low} \dot{x}_{ji} + \epsilon_i^{low}, \tag{2.15}$$

e

$$\ddot{y}_i = \beta_0^{upp} + \sum_{j=1}^{p} \beta_j^{upp} \ddot{x}_{ji} + \epsilon_i^{upp}, \tag{2.16}$$

respectivamente. Nessas expressões, $\epsilon_i^{low} (i = 1; \cdots; n)$ expressa os erros do modelo e $\beta_j^{low} (j = 1; \cdots; p)$ são os coeficientes do modelo. De forma similar, para o limite superior, $\epsilon_i^{upp} (i = 1; \cdots; n)$ e $\beta_j^{upp} (j = 1; \cdots; p)$ expressam os erros do modelo e os coeficientes, respectivamente.

Uma outra representação para a Equação 2.15, pode ser a forma matricial, reescrita como

$$y^{low} = X^{low} \beta^{low} + \epsilon^{low}, \tag{2.17}$$

em que

$$X^{low} = \begin{pmatrix} 1 & \dot{x}_{11} & \dot{x}_{21} & \cdots & \dot{x}_{p1} \\ 1 & \dot{x}_{12} & \dot{x}_{22} & \cdots & \dot{x}_{p2} \\ 1 & \vdots & \vdots & \cdots & \vdots \\ 1 & \dot{x}_{1n} & \dot{x}_{2n} & \cdots & \dot{x}_{pn} \end{pmatrix},$$

$$\beta^{low} = (\ \beta_0^{low} \quad \beta_1^{low} \quad \cdots \quad \beta_p^{low}\ )^T,$$

$$y^{low} = (\ \dot{y}_1 \quad \dot{y}_2 \quad \cdots \quad \dot{y}_n\ )^T,$$

$$\epsilon^{low} = (\ \epsilon_1^{low} \quad \epsilon_2^{low} \quad \cdots \quad \epsilon_n^{low}\ )^T.$$

A estimação por mínimos quadrados da Equação 2.17 é

$$\hat{\beta}^{low} = ((X^{low})^T X^{low})^{-1} (X^{low})^T y^{low} \tag{2.18}$$

De forma similar, a Equação 2.16 pode ser expressa em um formato matricial, dado por:

$$\ddot{y}_i = \beta_0^{upp} + \sum_{j=1}^{p} \beta_j^{upp} \ddot{x}_{ji} + \epsilon_i^{upp}, \tag{2.19}$$

em que

$$X^{upp} = \begin{pmatrix} 1 & \ddot{x}_{11} & \ddot{x}_{21} & \cdots & \ddot{x}_{p1} \\ 1 & \ddot{x}_{12} & \ddot{x}_{22} & \cdots & \ddot{x}_{p2} \\ 1 & \vdots & \vdots & \cdots & \vdots \\ 1 & \ddot{x}_{1n} & \ddot{x}_{2n} & \cdots & \ddot{x}_{pn} \end{pmatrix},$$

$$\beta^{upp} = (\ \beta_0^{upp} \quad \beta_1^{upp} \quad \cdots \quad \beta_p^{upp}\ )^T,$$

$$y^{upp} = (\ \ddot{y}_1 \quad \ddot{y}_2 \quad \cdots \quad \ddot{y}_n\ )^T,$$

$$\epsilon^{upp} = (\ \epsilon_1^{upp} \quad \epsilon_2^{upp} \quad \cdots \quad \epsilon_n^{upp}\ )^T.$$

A estimação por mínimos quadrados da Equação 2.19 é dada por:

$$\hat{\beta}^{upp} = ((X^{upp})^T X^{upp})^{-1} (X^{upp})^T y^{upp}. \tag{2.20}$$

Considere x_π o vetor de intervalos, a estimativa do limite inferior e o limite superior da resposta, são expressas por

$$\hat{\dot{y}}_\pi = x_\pi^{low} \hat{\beta}^{low}, \tag{2.21}$$

$$\hat{\ddot{y}}_\pi = x_\pi^{upp} \hat{\beta}^{upp}. \tag{2.22}$$

2.1.4.2 Método do Centro e Amplitude

O método do centro e amplitude apresentado nessa seção foi proposto por Neto e Carvalho (2008). A construção do modelo é feita baseada no centro e amplitude dos intervalos. O método do centro proposto por Billard e Diday (2006c), utiliza os centros dos intervalos, resposta e regressores, para construir o modelo de regressão linear. Os coeficientes do modelo construído e os limites dos regressores são usados na predição dos limites da resposta. O modelo de regressão linear proposto é especificado conforme

$$\begin{cases} \dot{y}_i = \beta_0^{cen} + \sum_{j=1}^{p} \beta_j^{cen} \dot{x}_{ji} + \epsilon_i^{low}, \\ \ddot{y}_i = \beta_0^{cen} + \sum_{j=1}^{p} \beta_j^{cen} \ddot{x}_{ji} + \epsilon_i^{upp}. \end{cases} \tag{2.23}$$

onde $\beta_j^{cen} (j = 0; \cdots; p)$ são os coeficientes da regressão; ϵ_i^{low} e ϵ_i^{upp} são os erros de estimação dos limites inferiores e superiores. Para o intervalo y_i, o erro de estimação do centro é definido

por $\epsilon_i^{cen} = (\epsilon_i^{low} + \epsilon_i^{upp})/2$. Para obter a estimação dos coeficientes utiliza-se o método dos mínimos quadrados, onde

$$\begin{aligned} S &= \sum_{i=1}^{n} (\epsilon_i^{cen})^2 \\ &= \sum_{i=1}^{n} \left(\frac{\dot{y}_i + \ddot{y}_i}{2} - \beta_0^{cen} - \sum_{j=1}^{p} \beta_j^{cen} \left(\frac{\dot{x}_{ji} + \ddot{x}_{ji}}{2} \right) \right)^2 \\ &= \sum_{i=1}^{n} \left(y_i^{cen} - \beta_0^{cen} - \sum_{j=1}^{p} \beta_j x_{ji}^{cen} \right)^2, \end{aligned} \tag{2.24}$$

em que $y_i^{cen} = \frac{(\dot{y}_i + \ddot{y}_i)}{2}$ e $x_{ji}^{cen} = \frac{(\dot{x}_{ji} + \ddot{x}_{ji})}{2}$, são os centros dos intervalos y_i e x_{ji}. A forma matricial da Equação 2.23 é representada por

$$y^{cen} = X^{cen}\beta^{cen} + \epsilon^{cen}, \tag{2.25}$$

onde

$$X^{cen} = \begin{pmatrix} 1 & x_{11}^c & x_{21}^c & \cdots & x_{p1}^c \\ 1 & x_{12}^c & x_{22}^c & \cdots & x_{p2}^c \\ 1 & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n}^c & x_{2n}^c & \cdots & x_{pn}^c \end{pmatrix},$$

$$\beta^{cen} = (\ \beta_0^{cen} \quad \beta_1^{cen} \quad \cdots \quad \beta_p^{cen}\ )^T,$$

$$y^{cen} = (\ \dot{y}_1 \quad \dot{y}_2 \quad \cdots \quad \dot{y}_n\ )^T,$$

$$\epsilon^{cen} = (\ \epsilon_1^{cen} \quad \epsilon_2^{cen} \quad \cdots \quad \epsilon_n^{cen}\ )^T.$$

Segundo Billard e Diday (2006c), a solução por mínimos quadrados é dada por

$$\hat{\beta}^{cen} = ((X^{cen})^T X^{cen})^{-1} (X^{cen})^T y^{cen}, \tag{2.26}$$

Para o cálculo da estimativa do limite inferior e superior da resposta para x_π, utiliza-se as seguintes equações

$$\hat{\dot{y}}_\pi = x_\pi^{low} \hat{\beta}^{cen}, \tag{2.27}$$

$$\hat{\ddot{y}}_\pi = x_\pi^{upp} \hat{\beta}^{cen}. \tag{2.28}$$

De maneira geral, a predição para o método do centro é simples e segue os mesmos preceitos utilizados na regressão para dados pontuais.

Uma outra abordagem é construída para a predição das amplitudes dos intervalos. Considere

$$\breve{y}_i = \beta_0^{amp} + \sum_{j=1}^{p} \beta_j^{amp} \breve{x}_{ji} + \epsilon_i^{amp}, \tag{2.29}$$

em que $\breve{y} = \dot{y}_i - \ddot{y}_i$, $\breve{x} = \dot{x}_{ij} - \ddot{x}_{ij}$, expressam as amplitudes dos intervalos y_i e x_{ji}; β_j^{amp} $(j = 0; \cdots; p)$ são os coeficientes da regressão e ϵ_i^{amp} $(i = 1; \cdots; n)$ são os erros do modelo. A forma matricial da Equação 2.29 é expressa pela seguinte equação:

$$y^{amp} = X^{amp}\beta^{amp} + \epsilon^{amp}, \tag{2.30}$$

em que

$$X^{amp} = \begin{pmatrix} 1 & \breve{x}_{11} & \breve{x}_{21} & \cdots & \breve{x}_{p1} \\ 1 & \breve{x}_{12} & \breve{x}_{22} & \cdots & \breve{x}_{p2} \\ 1 & \vdots & \vdots & \cdots & \vdots \\ 1 & \breve{x}_{1n} & \breve{x}_{2n} & \cdots & \breve{x}_{pn} \end{pmatrix},$$

$$\beta^{amp} = (\ \beta_0^{amp} \quad \beta_1^{amp} \quad \cdots \quad \beta_p^{amp}\ )^T,$$

$$y^{amp} = (\ \breve{y}_1 \quad \breve{y}_2 \quad \cdots \quad \breve{y}_n\ )^T \text{e}$$

$$\epsilon^{amp} = (\ \epsilon_1^{amp} \quad \epsilon_2^{amp} \quad \cdots \quad \epsilon_n^{amp}\ )^T.$$

Utiliza-se o método dos mínimos quadrados para a estimativa dos coeficientes (Equação 2.31). Para o cálculo da predição usando x_π, considere $x_\pi^{cen} = (1\ x_{\pi 1}^{cen}\ x_{\pi 2}^{cen}\ \cdots\ x_{\pi p}^{cen})$ e $\breve{x}_\pi^{amp} = (1\ \breve{x}_{\pi 1}\ \breve{x}_{\pi 2}\ \cdots\ \breve{x}_{\pi p})$, com $x_{\pi j}^{cen} = (\ddot{x}_{\pi j} + \dot{x}_{\pi j})/2$ e $\breve{x}_{\pi\ j} = (\ddot{x}_{\pi j} - \dot{x}_{\pi j})$.

$$\hat{\beta}^{amp} = ((X^{amp})^T X^{amp})^{-1} (X^{amp})^T y^{amp}. \tag{2.31}$$

Para obtenção da estimativa do centro utiliza-se a Equação 2.32. De forma similar, a amplitude da resposta é predita pela Equação 2.33

$$\hat{y}_\pi^{cen} = x_\pi^{cen} \hat{\beta}^{cen}, \tag{2.32}$$

$$\hat{\breve{y}}_\pi = x_\pi^{amp} \hat{\beta}^{amp}. \tag{2.33}$$

Os limites de predição são determinados respectivamente pelas Equações 2.34 e 2.35.

$$\hat{\ddot{y}}_\pi = \hat{y}_\pi^c - \frac{\hat{\breve{y}}_\pi}{2}, \tag{2.34}$$

$$\hat{\hat{y}}_\pi = \hat{y}_\pi^c + \frac{\hat{\hat{y}}_\pi}{2}. \tag{2.35}$$

De acordo com Neto e Carvalho (2008), o modelo centro-amplitude apresenta um melhor desempenho, visto que o mesmo consegue melhorar a predição dos intervalos, quando existe uma dependência linear entre as amplitudes dos regressores e da resposta.

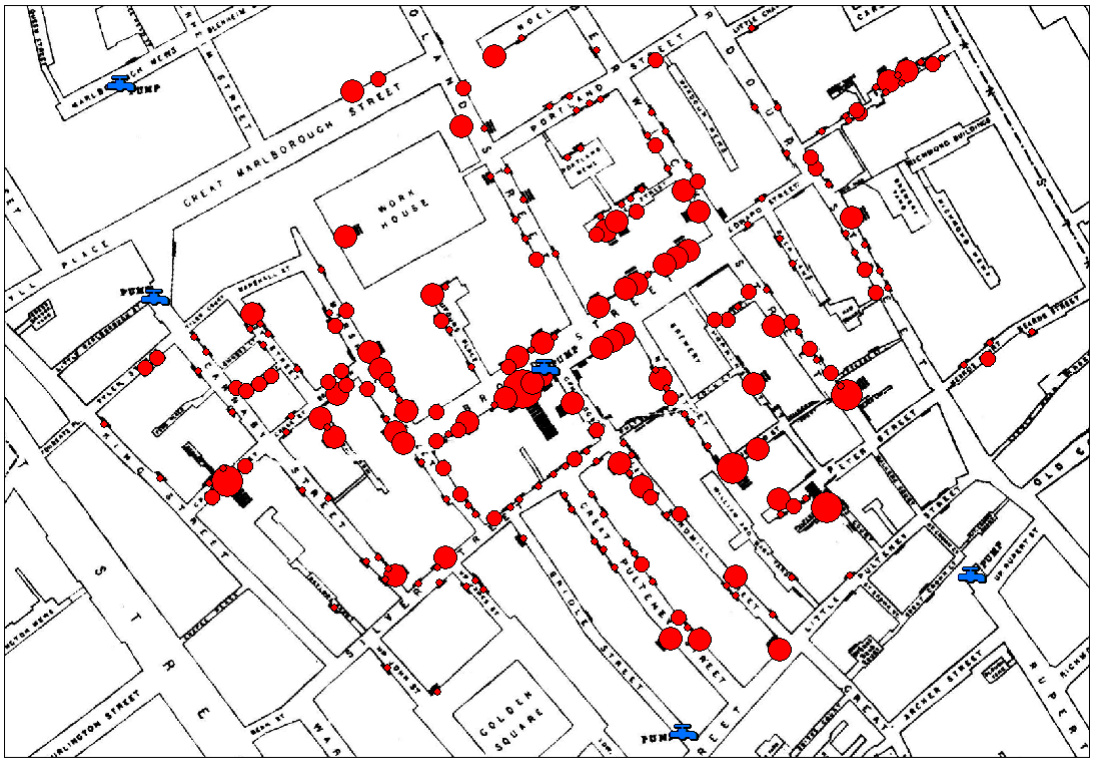
2.2 ANÁLISE DE DADOS ESPACIAIS

O termo estatística espacial é utilizado para descrever uma ampla gama de modelos e métodos estatísticos destinados à análise de dados referenciados espacialmente. De acordo com Cressie (1992), este tipo de dado pode ser visto como resultado de um processo estocástico, e estatísticas mais realistas podem ser encontradas. Bivand et al. (2008) fornecem uma visão geral dessas metodologias.

Compreender a distribuição espacial de dados resultantes de fenômenos no espaço tem se tornado um desafio para diversas áreas do conhecimento. A análise espacial fornece métodos de investigação desses fenômenos, que apresentam uma dependência espacial que não pode ser ignorada, como, por exemplo, fenômenos relacionados à saúde, educação, tecnologia, economia, entre outros.

As primeiras manifestações de estatísticas para dados espaciais parecem ter surgido na forma de dados de mapas. No campo da epidemiologia, por exemplo, a primeira análise incorporando a noção de espaço foi realizada no século XIX por John Snow (BIVAND et al., 2008). Snow demonstrou que a cólera era causada pelo consumo de águas contaminadas com matérias fecais em um poço localizado na Broad Street, ao comprovar que os casos dessa doença se agrupavam nas zonas onde a água consumida estava contaminada com fezes, na cidade de Londres no ano de 1854. Mais detalhes sobre esse estudo podem ser obtidos em (SNOW; RICHARDSON; FROST, 1936).

Figura 3 – Mapa de pontos de Snow das mortes por cólera em 1854, Golden Square.



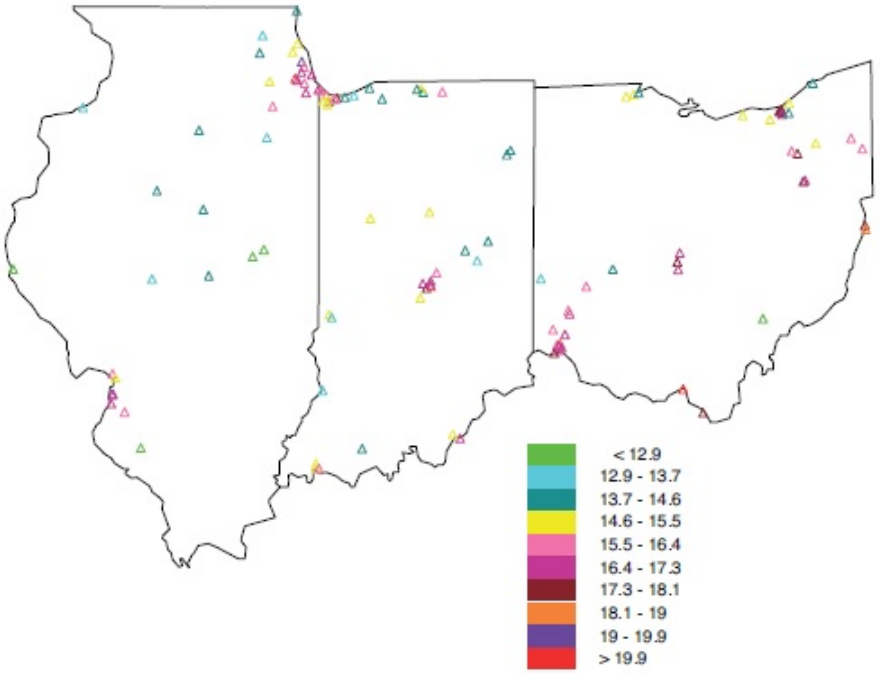
Fonte: Figura adaptada de Câmara et al. (2004).

A análise espacial é capaz de mensurar propriedades e relacionamentos, levando em conta a localização espacial do fenômeno em estudo. Nesse contexto, o tipo de dado analisado incopora informações sobre o espaço no qual esta inserido. Segundo Bivand et al. (2008), os tipos de dados mais utlizados para caracterizar os problemas de análise espacial consideram as seguintes abordagens:

- **Eventos ou Padrões Pontuais** - Fenômenos expressos através de ocorrências identificadas como pontos localizados no espaço denominados processos pontuais;
- **Superfícies contínuas** - Estimadas a partir de uma amostra de campo, que podem estar regularmente ou irregularmente distribuídas;
- **Áreas com contagens e Taxas agregadas** - Tratam-se de dados associados a levantamentos populacionais, como censos e estatísticas de sáude, e que geralmente se referem a indivíduos localizados em pontos específicos do espaço.

Para visualização desses tipos de dados, considere as Figuras 4-6. A Figura 4 apresenta um exemplo de dados pontuais, mostrando as localizações de 114 locais de monitoramento

Figura 4 – Exemplo de eventos padrões com a localização de monitoramento da poluição do ar em três estados dos EUA.



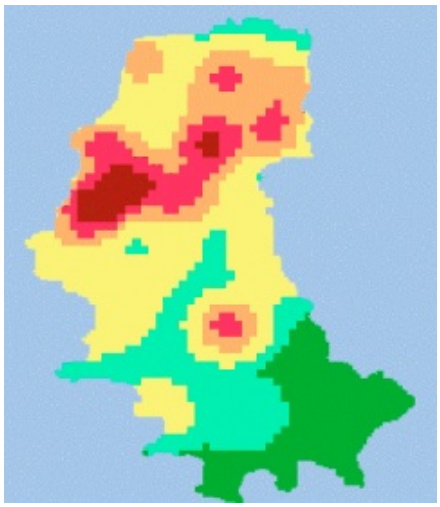
Fonte: Figura adaptada de Câmara et al. (2004).

da poluição do ar em três estados do meio-oeste dos Estados Unidos (Illinois, Indiana e Ohio) (BANERJEE; CARLIN; GELFAND, 2014).

Como exemplo para dados de superfícies contínuas, considere a cidade de Porto Alegre, mostrada na Figura 5. Esse mapa ilustra uma superfície para intensidade estimada, que pode ser pensada como a 'temperatura' da violência por causas externas. A superfície interpolada mostra um padrão de distribuição de pontos com uma forte concentração no centro da cidade e decrescendo aos bairros mais afastados.

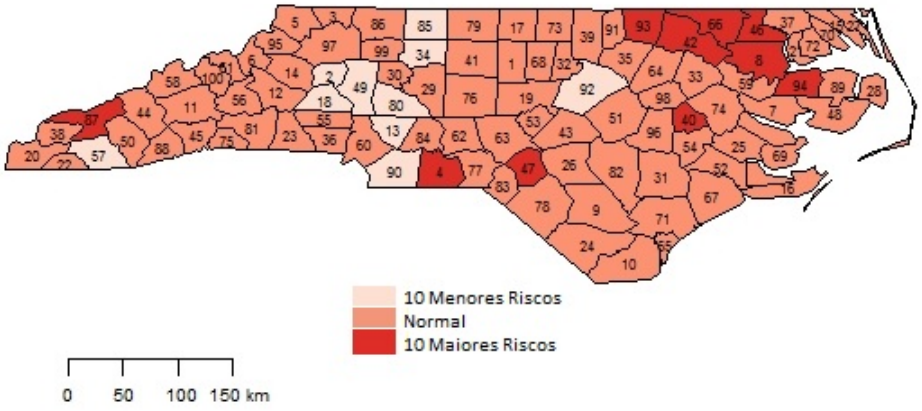
Na prática, a maioria dos dados de área são resumos sobre uma rede irregular, como uma coleção de condados ou outras fronteiras regionais. Como exemplo, considere algumas informações sobre o conjunto de dados *nc.sids* para os condados da Carolina do Norte do pacote *spData* do R (CRESSIE, 1992). A superfície mostra a distribuição espacial do risco de mortalidade infantil relacionados à uma síndrome. Na Figura 6, em que cada região é descrita pelo seu número de identificação, estão em destaque os condados que apresentam os 10 mais altos e os 10 mais baixos valores para essa taxa. Essa figura é um exemplo de mapa coroplético, o que significa que ele usa tons de cor (ou escala de cinza) para classificar os valores em algumas classes amplas.

Figura 5 – Exemplo de eventos contínuos com a intensidade da violência por causas externas em Porto Alegre.



Fonte: Figura adaptada de Câmara et al. (2004).

Figura 6 – Exemplo com taxas agregadas do risco por morte súbita infantil para o período de 1974-1978, na Carolina do Norte.



Fonte: Figura adaptada de Cressie (1992).

De acordo com as abordagens citadas acima, encontram-se problemas de análise espacial em diversas áreas da literatura. Para tratamento desses dados, o processo de modelagem é precedido de uma fase de análise exploratória, associada à apresentação visual dos dados sob forma de gráficos, mapas e a identificação de padrões de dependência espacial do fenômeno em estudo.

Um conceito importante na compreensão e análise dos fenômenos espaciais é a dependência espacial. Essa noção parte de Tobler (1970), no qual afirma que a maior parte das ocorrências, sejam estas naturais ou sociais, apresentam entre si uma relação que dependem da distância.

Desde então, várias teorias e métodos analíticos relacionados ao estudo de padrões espaciais de incidência de doenças e mortalidade foram propostos. Neste contexto, os mapas são utilizados como ferramentas úteis para indicar comportamentos relacionados com a origem e propagação de doenças (WALLER; GOTWAY, 2004). Um outro aspecto relevante neste contexto se refere ao estudo da autocorrelação espacial, com o intuito de descrever o comportamento dos dados georreferenciados. Os dados são georreferenciados quando a localização da observação é expressa por um par de coordenadas geográficas (HAINING; HAINING, 2003). As coordenadas também podem ser representadas pelo centroide de uma região, pois é uma forma representativa de selecionar um ponto em uma determinada área. Por exemplo, Bivand e Wong (2018) compararam as implementações de indicadores globais e locais de associação espacial utilizando os centroides das regiões.

A análise de autocorrelação espacial inclui uma classe de métodos que medem a dependência espacial, a associação entre o valor em um local específico e os valores para áreas próximas ou adjacentes (CRESSIE, 1992). Mais formalmente, a existência de autocorrelação espacial pode ser expressa por

$$Cov(y_i, y_j) = E(y_i, y_j) - E(y_i)E(y_j) \ , \ i \neq j, \tag{2.36}$$

onde y_i e y_j são observações de uma variável aleatória nas localizações i e j no espaço. Tais localizações podem ser pontos (por exemplo, áreas metropolitanas, medidas como latitude e longitude) ou unidades de área (por exemplo, estados). A autocorrelação espacial significa que o valor de uma variável de interesse em uma determinada região i está associado ao valor dessa variável na região vizinha j.

A autocorrelação espacial mede a correlação de uma variável consigo mesma quando as observações são consideradas no espaço. É definida como uma correlação positiva ou negativa de uma variável consigo mesma devido à localização espacial das observações (Haining e Haining (2003)). Em geral, os índices de autocorrelação espacial são usados para caracterizar a correlação entre medidas geograficamente semelhantes para um mesmo fenômeno observado. Quando as observações em um determinado local são fortemente influenciadas por sua vizinhança, a correlação espacial correspondente será alta. Esta medida visa identificar se um dado fenômeno ocorre de forma aleatória ou se existe um padrão espacial (Anselin (1995), Bivand et al. (2008), Matheron (1963), Getis e Boots (1978)). Para verificar essa premissa, podemos encontrar diversas metodologias na literatura. Uma das mais utilizadas é o índice de Moran, que abordaremos a seguir.

2.2.1 Índice de Autocorrelação Espacial

A Análise exploratória de dados espaciais permite identificar padrões espaciais não esperados, buscando identificar áreas de alta prevalência de determinados fenômenos e formular hipóteses para orientar pesquisas futuras. Segundo Bailey e Gatrell (1995), existem três tipos gerais de análise espacial: visualização, análise exploratória de dados e construção de modelos.

Na análise exploratória de dados espaciais, podemos verificar padrões espaciais observando a distribuição dos pontos em uma região. Diversas medidas de autocorrelação espacial têm sido propostas para investigar o processo espacial da evolução geográfica de diferentes pontos de vista. Dentre os principais índices utilizados para medir a autocorrelação espacial destacam-se o Índice de Moran (I), o índice de Geary e a estatística de Getis-Ord. Além disso, o teste de Tango pode ser utilizado para verificar a existência ou não de padrões espaciais.

Segundo Goodchild (1986), o índice de Geary assume que a interação não é o produto cruzado dos desvios da média, mas sim as intensidades dos desvios de cada local de observação com outro. Uma aplicação desse índice na literatura pode ser encontrada em Braz et al. (2009). Já a estatística de Getis-Ord, é proposta considerando duas abordagens para estimar a autocorrelação espacial, Getis e Ord (1992) e Ord e Getis (1995). Esses autores definiram duas estatísticas, G_i e G_i^*, que diferem no fato da localidade central i estar incluída no cálculo do coeficiente. Essa estátistica pode ser encontrada em Sokal e Thomson (2006).

Tango (1995) propôs um teste semelhante de agrupamento global, no qual o número de casos observados e esperados em cada região são comparados. Ele ressalta que diferentes tipos de interações existentes entre regiões vizinhas podem ser considerados e propõe uma medida de relação baseada em uma função da distância entre duas regiões.

Dentre os índices apresentados, o mais comumente utilizado é o índice de Moran, que é uma medida de autocorrelação espacial desenvolvida por Patrick Alfred Pierce Moran (MORAN, 1950). Esse índice é utilizado para verificar dependências espaciais, no sentido de indicar as circunstâncias em que o valor da observação de uma única localização depende do valor das observações vizinhas, buscando identificar padrões e níveis de agrupamento espacial.

Metodologias para o cálculo do Índice de Moran podem ser encontradas em Chen (2013). De acordo com a literatura, o índice clássico de autocorrelação espacial de Moran tem encontrado ampla aplicação em diversas áreas do conhecimento, como epidemiologia (ER et al., 2010); ecologia (DORMANN et al., 2007) e economia (BASU; THIBODEAU, 1998). Nesse contexto, indicadores locais de estatísticas de autocorrelação espacial (ANSELIN, 1995) e um estudo de

clusters espaciais e *outliers* (TSAI; PERNG, 2011) foram reportados.

Seja X uma variável clássica contínua e Ω um conjunto de dados de tamanho n. O índice global de Moran para dados clássicos é medido por meio da seguinte equação:

$$I = \frac{n}{\sum_i^n \sum_j^n w_{ij}} \frac{\sum_i^n \sum_j^n w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i^n (x_i - \bar{x})^2}, \tag{2.37}$$

onde n é o número de observações indexadas para i e j; $\bar{x}$ é o valor médio de X; x_i e x_j são os valores das observações nas regiões i e j, respectivamente; w_{ij} é o elemento da matriz de proximidade **W**, elemento $(i = 1, \ldots, n)$ $(j = 1, \ldots, n)$, a qual expressa a relação espacial entre as n populações.

Esse índice fornece uma única medida para o conjunto de todas as áreas, caracterizando toda a região de estudo. Os valores do índice variam entre –1 a 1. O grau de autocorrelação é positivo para correlação direta, negativo quando inversa, e nulo ou com valores próximos de zero quando os indicadores se distribuem ao acaso entre as áreas vizinhas, sem relação espacial.

A significância do índice pode ser avaliada sob a hipótese nula, por duas suposições básicas: normalidade ou aleatoriedade. Na primeira, os dados são observações independentes provenientes de uma população com distribuição normal e na segunda, a função de distribuição é desconhecida (LOUZADA, 2006).

Na literatura, várias métricas são consideradas para se obter a matriz **W** com base em um critério de vizinhança. Dado um amplo escopo de matrizes de pesos espaciais, o cálculo da matriz de pesos é um processo importante, pois fornece o grau de conectividade entre as regiões. A matriz de pesos é utilizada para descrever a estrutura de dependência espacial entre as unidades de análise, sendo os pesos definidos utilizando diferentes métricas. Todas as matrizes consideradas são binárias, entretanto, não há restrição para o uso de matrizes não binárias.

A estrutura de vizinhança de um conjunto de dados é uma forma de representar a interação entre unidades espaciais. Esta representação é dada por uma matriz quadrada **W**, de dimensão $n \times n$, onde n é o número de regiões consideradas. Os pesos w_{ij} são calculados de acordo com algum critério de proximidade, mostrando a influência da região j sobre a região i. Assim, a matriz de pesos **W** é utilizada como ponderação que busca identificar a influência que as regiões exercem umas sobre as outras (GOODCHILD, 1986).

As matrizes de ponderação espacial podem ser classificadas de acordo com um critério

geográfico ou socioeconômico (FISCHER; WANG, 2011). No critério geográfico, cada área é representada por um vetor de centroides onde as coordenadas cartesianas são conhecidas. A matriz de ponderação pode ser definida de acordo com a contiguidade e/ou distância geográfica utilizando uma determinada métrica. Dentre os vários critérios propostos na literatura, apresentaremos uma breve revisão dos utilizados neste trabalho, os quais são expressos por: critério de contiguidade da rainha (FISCHER; WANG, 2011), K-centróides mais próximos (kcn) com $K = 1{,}2$ (NEGREIROS et al., 2010) e distância de Chebyshev.

A matriz de ponderação baseada na ideia de contiguidade é uma matriz binária em que as regiões vizinhas compartilham uma fronteira física comum, de forma que regiões que compartilham pelo menos um ponto na fronteira, ou seja, tendo latitudes e longitudes comuns, são consideradas regiões vizinhas. A escolha dos critérios de contiguidade resulta em estruturas vizinhas muito diferentes. Neste estudo, utilizamos o critério de contiguidade da rainha, no qual todas as regiões que compartilham uma fronteira comum são consideradas vizinhas. O peso é definido como:

$$w_{ij} = \begin{cases} 1, \text{ se } R_i \text{ e } R_j \text{ são contíguos;} \\ 0, \text{ Caso contrário.} \end{cases} \tag{2.38}$$

em que R_i e R_j são áreas que apresentam bordas em comum.

Com base no critério do kcn com $K = 1{,}2$, **W** é uma matriz binária computada de acordo com o procedimento apresentado em Negreiros et al. (2010). Uma vez que cada região é representada por um vetor de centroides, a ideia é encontrar K-centroides mais próximos desta região usando a distância euclideana entre dois vetores de centroides. Dados os vetores de centroides m_i e m_j das regiões R_i e R_j, respectivamente, o peso é dado por

$$w_{ij} = \begin{cases} 1, \text{ se } m_j \text{ é um dos } K\text{-centroides mais próximos } m_i; \\ 0, \text{ caso contrário.} \end{cases} \tag{2.39}$$

Se empregarmos o critério de distância (FISCHER; WANG, 2011) para calcular a matriz de peso binário **W**, a vizinhança de uma região R_i é definida pela distância de Chebyshev entre os vetores de centroides m_i e m_j e comparado com um valor d, conforme segue,

$$w_{ij} = \begin{cases} 1, \text{ se } d(m_i, m_j) < \text{d;} \\ 0, \text{ caso contrário.} \end{cases} \tag{2.40}$$

em que d é o segundo menor valor por linha da matriz de vizinhança.

A autocorrelação espacial positiva é representada pela similaridade entre as populações (polígonos). Ou seja, as regiões que possuem um alto/baixo valor para a variável observada tendem a ser vizinhos de populações (polígonos) que também possuem um alto/baixo valor. Em contrapartida, a autocorrelação espacial negativa indica que existe uma dissimilaridade entre os valores observados e de sua localização espacial.

2.2.2 Modelos de Regressão para Dados Espaciais

A ausência de independência entre as observações é uma característica inerente dos dados espaciais. Em um mundo ideal, seria preferível coletar dados nos quais as observações fossem mutuamente independentes e, assim, evitar problemas na inferência dos resultados.

Devido a essa característica, ajustar modelos na presença de autocorrelação espacial tornou-se um desafio, uma vez que devemos considerar que as observações não são independentes, porque pode haver alguma correlação entre as áreas vizinhas.

De acordo com a literatura, a dependência espacial pode ser modelada de diferentes maneiras utilizando modelos estatísticos. Cressie (1992) fornece uma discussão dessas abordagens, e alguns exemplos são apresentados. Modelos espaciais autoregressivos são apresentados em Schabenberger e Gotway (2017). Wall (2000) fornece um estudo comparativo das maneiras pelas quais os processos espaciais são modelados. Dentre as abordagens mais utilizadas, destacam-se os modelos Autoregressivo Simultâneo (SAR) e o Autoregressivo Condicionalmente (CAR).

Segundo Bivand et al. (2008), a especificação SAR utiliza uma regressão nos valores de outras áreas para contabilizar a dependência espacial. Isso significa que os termos de erro e_i são modelados de modo que dependam uns dos outros. Já a especificação do modelo CAR depende da distribuição condicional dos termos de erro espacial.

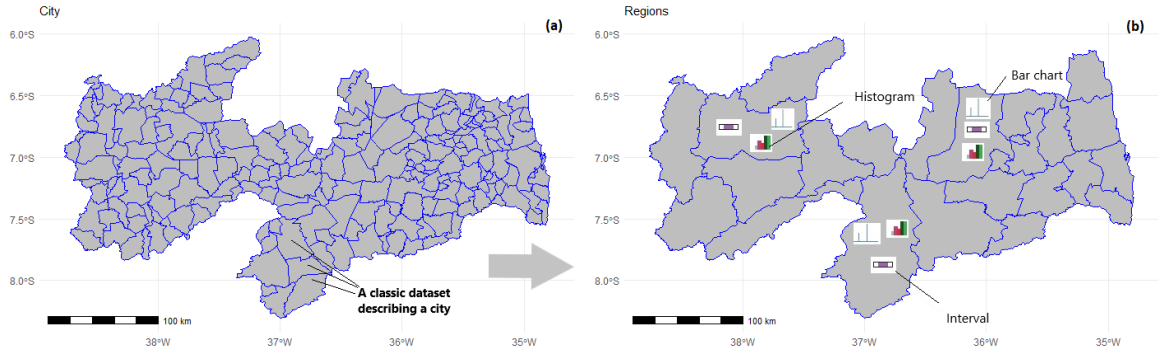
3 ÍNDICES DE AUTOCORRELAÇÃO ESPACIAL INTERVALAR

Várias metodologias da análise espacial são propostas na literatura para investigar o comportamento de dados georreferenciados. Para investigar a existência ou não de padrões espaciais encontramos diversas referências na literatura (GOODCHILD, 1986; GETIS; ORD, 1992; ORD; GETIS, 1995; MORAN, 1950). No contexto de modelagem, podemos encontrar modelos de regressão que levam em consideração a dependência espacial existente (BIVAND et al., 2008; SCHABENBERGER; GOTWAY, 2017). Dentre as diversas abordagens, podemos utilizar índices de autocorrelação espacial, modelos de regressão espacial, risco relativo, entre outras.

Este capítulo apresenta uma generalização de um índice de autocorrelação espacial, onde a variação da informação contida entre os limites dos intervalos é levada em consideração. O índice proposto é observado na presença e ausência de autocorrelação espacial, sob diferentes parametrizações. Além disso, consideramos diferentes abordagens para o cálculo da matriz de pesos, que é utilizada para a identificação da região de vizinhança.

Segundo Lima et al. (2019), unidades geográficas representam objetos geográficos (ou geo-objetos). Para nos ajudar a visualizar como os dados simbólicos estão relacionados ao espaço geográfico, considere o exemplo relacionado a um conjunto de cidades na Figura 7. Nesta figura, um conjunto de n geo-objetos de vizinhança (Figura 7(a)), formam um novo geo-objeto (Figura 7(b))), com $n \geq 1$. Cada unidade é inicialmente descrita por dados clássicos como valores categóricos ou contínuos (lado esquerdo). Em seguida, as cidades são agregadas por regiões para obter um novo conjunto de dados (lado direito) em que as novas unidades são classes de cidades individuais. Essas novas unidades podem ser descritas por variáveis simbólicas como um histograma, um gráfico de barras de categorias, intervalos ou uma lista de valores que expressam a variabilidade das cidades dentro de cada região em vez de valores únicos, como por exemplo, centroides de áreas.

Figura 7 – Representação de um mapa com dados padrões agregados para um mapa com dados simbólicos.



Fonte: Elaborada pela autora.

3.1 ÍNDICES INTERVALARES

Com a finalidade de estudar um índice de autocorrelação espacial para variáveis no contexto simbólico, generalizamos um índice clássico de autocorrelação espacial, pois em algumas situações os dados são fornecidos agregados. Mais especificamente, iremos apresentar uma proposta de generalização do Índice de Moran para dados simbólicos (FREITAS et al., 2022). Nesse contexto, é proposto um índice que relaciona a área de análise espacial com a análise de dados simbólicos. Esse índice é capaz de mensurar a autocorrelação espacial existente levando em cosideração a informação georeferenciada e uma varíavel simbólica.

Para isso, duas representações são consideradas para o cálculo desta versão: mínimo-máximo e centro-amplitude. Dessa maneira, esses índices são dados pelas definições a seguir.

Definição 6 *O Índice de Autocorrelação Simbólico intervalar com base na representação mínimo-máximo (IASi) é dado por*

$$IASi = \frac{m}{\sum_k^n \sum_t^n \lambda_{kt}} \frac{\sum_k^n \sum_t^n \lambda_{kt} \left[(a_k - \bar{a})(a_t - \bar{a}) + (b_k - \bar{b})(b_t - \bar{b})\right]}{\sum_k^n \left[(a_k - \bar{a})^2 + (b_k - \bar{b})^2\right]}, \tag{3.1}$$

onde λ_{kt} é um peso associado às posições espaciais de R_k e R_t, m é o número de regiões espaciais consideradas e $\bar{a}$ e $\bar{b}$ são dados por

$$\bar{a} = \frac{1}{m} \sum_{k \in E} a_k, \quad \bar{b} = \frac{1}{m} \sum_{k \in E} b_k. \tag{3.2}$$

Proposição 1 *O índice de autocorrelação espacial intervalar da Equação (3.1) é uma extensão do índice de Moran na Equação (2.37). Os valores encontrados para este índice estão entre*

-1 *e* $+1$ *. O valor* $+1$ *corresponde a uma autocorrelação forte, enquanto* -1 *indica uma autocorrelação espacial fraca. Um valor próximo a* 0 *indica um padrão espacialmente aleatório.*

Prova 1 *Assumindo intervalos degenerados, isto é, limites inferiores são iguais aos limites superiores (*$a_i = b_i$*), temos*

$$\begin{aligned} IASi &= \frac{m}{\sum_k^n \sum_t^n \lambda_{kt}} \frac{\sum_k^n \sum_t^n \lambda_{kt} \left[(a_k - \bar{a})(a_t - \bar{a}) + (a_k - \bar{a})(a_t - \bar{a})\right]}{\sum_k^n \left[(a_k - \bar{a})^2 + (a_k - \bar{a})^2\right]} \\ &= \frac{m}{\sum_k^n \sum_t^n \lambda_{kt}} \frac{2\sum_k^n \sum_t^n \lambda_{kt} \left[(a_k - \bar{a})(a_t - \bar{a})\right]}{2\sum_k^n (a_k - \bar{a})^2} \\ &= \frac{m}{\sum_k^n \sum_t^n \lambda_{kt}} \frac{\sum_k^n \sum_t^n \lambda_{kt} \left[(a_k - \bar{a})(a_t - \bar{a})\right]}{\sum_k^n (a_k - \bar{a})^2} \end{aligned}$$

Fixado o peso λ_{kt}*, podemos observar que o termo* $(x_i - \bar{x})(x_j - \bar{x})$ *na Equação (2.37) é substituído pela expressão* $[(a_k - \bar{a})(a_t - \bar{a}) + (b_k - \bar{b})(b_t - \bar{b})]$ *na Equação (3.1), e a variabilidade do intervalo é considerada. Ideia similar pode ser aplicada ao denominador. Semelhante ao índice clássico, os valores obtidos neste índice pertencem ao intervalo* $[-1; +1]$*, cuja interpretação é análoga ao índice de Moran.*

Considere que cada intervalo pode ser representado pelo par (c_i, r_i) onde $c_i = (b_i + a_i)/2$ e $r_i = (b_i - a_i)$ são o centro e amplitude do intervalo z_i.

Definição 7 *O Índice de Autocorrelação Simbólico intervalar com base na representação centro-amplitude (IASi*) é dado por*

$$IASi^* = \frac{m}{\sum_k^n \sum_t^n \lambda_{kt}} \frac{\sum_k^n \sum_t^n \lambda_{kt} \left[(c_k - \bar{c})(c_t - \bar{c}) + (r_k - \bar{r})(r_t - \bar{r})\right]}{\sum_k^n \left[(c_k - \bar{c})^2 + (r_k - \bar{r})^2\right]}, \tag{3.3}$$

onde λ_{kt} *é um peso associado às posições espaciais de* R_k *e* R_t *e* $\bar{c}$ *e* $\bar{r}$ *são definidos como, respectivamente,*

$$\bar{c} = \frac{1}{m} \sum_{k \in E} c_k \quad \textit{e} \quad \bar{r} = \frac{1}{m} \sum_{k \in E} r_k. \tag{3.4}$$

Proposição 2 *O índice de autocorrelação espacial de intervalo* $IASi^*$ *em relação à representação de centro-amplitude na Equação (3.3) é uma extensão do índice de Moran na Equação (2.37) para dados de intervalo e tem um valor entre* -1 *e 1. O valor* $+1$ *corresponde a uma autocorrelação forte, enquanto* -1 *indica uma autocorrelação fraca. Um valor próximo a* 0 *indica um padrão espacialmente aleatório.*

Prova 2 *Assumindo intervalos degenerados, ou seja, que o limite inferior e superior dos intervalos são iguais, a prova é semelhante à obtida na Proposição 1.*

3.1.1 Cálculo do peso binário

Agora, vamos considerar o caso em que a matriz de peso $\mathbf{\Lambda}$ é calculada com base no critério do K-vizinhos mais próximos (knn), utilizando uma distância intervalar. Utilizamos pesos binários para este cálculo. Cada região R é identificada como um intervalo bivariado que descreve a variabilidade das coordenadas de latitude e longitude dentro de R. Seja $([l_a, l_b], [g_a, g_b])$ um intervalo bivariado onde l indica a latitude e g representa a longitude. Os subscritos a e b indicam os valores inferior e superior para cada intervalo. Esses valores estão associados à região R.

Agora, vamos discutir brevemente como obtemos a matriz de pesos. Para o cálculo da matriz binária, cada região é representada por um intervalo bivariado que descreve a latitude e longitude. O intervalo bivariado pode ser pensado como um vetor de intervalos, que contém informação sobre os valores mínimo e máximo para a latitude e para a longitude. A ideia é encontrar K-vizinhos mais próximos desta região usando uma distância entre dois vetores de intervalo da literatura de SDA. Dados os vetores $\mathbf{v}_k = ([l_{ak}, l_{bk}], [g_{ak}, g_{bk}])$ e $\mathbf{v}_t = [l_{at}, l_{bt}], [g_{at}, g_{bt}])$ das regiões R_k e R_t, respectivamente, o peso é dado por

$$\lambda_{kt} = \begin{cases} 1, \text{ se } R_k \text{ é um dos } K\text{-vizinhos mais próximos de } R_t; \\ 0, \text{ caso contário.} \end{cases} \tag{3.5}$$

Para calcular a matriz de peso $\mathbf{\Lambda}$, utilizamos as seguintes distâncias: 2.5, 2.12, 2.13 e 2.14.

3.1.2 Validação Bootstrap

A abordagem de bootstrap foi introduzida por Efron (1979) para obter o intervalo bootstrap percentil dos índices de autocorrelação espacial. Segundo Rizzo (2019), é baseada na ideia de considerar uma certa amostra observada como sendo uma população finita, logo utiliza-se a reamostragem para estimar a distribuição da verdadeira população.

Para obter o intervalo bootstrap percentil, utiliza-se as B réplicas bootstrap de $\hat{\theta}$ para obter a função de distribuição empírica (fde) de $\hat{\theta}$ que denotaremos por $\hat{F}$. Dessa maneira, a partir dos percentis $(\frac{\alpha}{2})$ e $\left(1 - \frac{\alpha}{2}\right)$ de $\hat{F}$, obtem-se o intervalo percentil bootstrap com confiança de $(1 - \alpha) \times 100\%$, expresso por

$$\left[\hat{F}^{-1}\left(\frac{\alpha}{2}\right), \hat{F}^{-1}\left(1 - \frac{\alpha}{2}\right)\right],$$

em que $\hat{F}^{-1}\left(\frac{\alpha}{2}\right) = \hat{\theta}^{*(\frac{\alpha}{2})}$ e $\hat{F}^{-1}\left(1-\frac{\alpha}{2}\right) = \hat{\theta}^{*(1-\frac{\alpha}{2})}$, sendo reescrito da seguinte forma:

$$\left[\hat{\theta}^{*(\alpha/2)},\, \hat{\theta}^{*(1-\frac{\alpha}{2})}\right]. \tag{3.6}$$

Segundo Rizzo (2019), pela forma da construção deste intervalo, não é possível a presença de valores impróprios para o parâmetro θ no intervalo.

A metodologia da construção do intervalo percentil incorpora a transformação que normaliza a distribuição de $\hat{\theta}$ (EFRON, 1992). Suponha que $\hat{\theta}$ não possui distribuição normal em amostras finitas, e seja $\eta = h(\hat{\theta})$ a transformação que normaliza a distribuição de $\hat{\theta}$, ou seja,

$$\eta \sim \mathcal{N}(\eta, \tau^2), \tag{3.7}$$

onde τ é um desvio-padrão constante. Desta maneira, pode-se construir um intervalo de confiança exato com nível de confiança $(1-\alpha)\times 100\%$, dado por

$$\left[\hat{\eta} - z_{(1-\frac{\alpha}{2})}\tau;\, \hat{\eta} + z_{(1-\frac{\alpha}{2})}\tau\right], \tag{3.8}$$

em que $z_{(1-\frac{\alpha}{2})}$ é obtido de $P(-z_{(1-\frac{\alpha}{2})}) < Z < P(z_{(1-\frac{\alpha}{2})})) = 1-\alpha$ com $Z \sim \mathcal{N}(0,1)$. Assim, o intervalo percentil pode ser reescrito da forma

$$\left[h^{-1}(\hat{\eta} - z_{(1-\frac{\alpha}{2})}\tau);\, h^{-1}(\hat{\eta} + z_{(1-\frac{\alpha}{2})}\tau)\right],$$

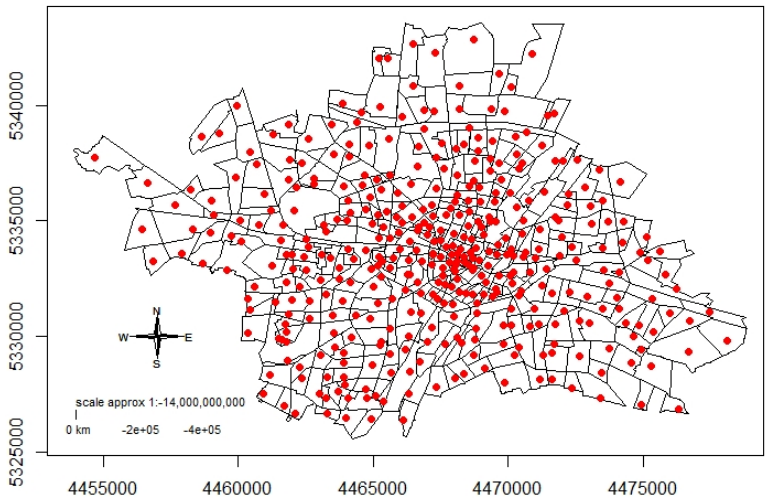
sem a necessidade de conhecer a função transformadora $h(\cdot)$.

3.2 AVALIAÇÃO EXPERIMENTAL

Para mostrar a utilidade da abordagem proposta neste trabalho, experimentos sintéticos com conjuntos de dados de intervalo relativos a diferentes configurações são considerados nesta seção. Em seguida, é realizada uma análise buscando identificar qualquer padrão existente em diferentes cenários. Os conjuntos de dados são gerados com a presença e ausência de autocorrelação espacial para uma determinada região.

Segundo Cressie (1992), uma vez que o tipo de área é irregular, as análises obtidas são mais realistas, visto que as regiões apresentam comportamentos semelhantes com a realidade. Com essa finalidade, o mapa da cidade de *Munique* (Alemanha) é utilizado (Figura 8) para o estudo de simulação dos dados. Esse mapa tem como característica a existência de regiões irregulares e informações em sua base de dados que são relevantes para o estudo.

Figura 8 – Mapa da cidade de Munique com os centroides por regiões.



Fonte: Elaborada pela autora.

Nesta seção, iremos apresentar um estudo comparativo entre as representações mínimo-máximo e centro-amplitude. Também consideramos o caso para a amplitude intervalar nula. Além disso, conjuntos de dados sintéticos clássicos também são considerados neste estudo com o objetivo de mostrar a importância de se levar em consideração a variabilidade dentro das regiões no cálculo da autocorrelação espacial. Em seguida, os índices de autocorrelação espacial para conjuntos de dados clássicos e de intervalo são obtidos e analisados neste estudo.

Inicialmente, dados clássicos sintéticos são gerados seguindo uma distribuição normal para todas as 411 regiões da cidade de *Munique* e diferentes valores de parâmetros da distribuição são adotados. Os dados intervalares são obtidos para cada região (classe) a partir dos valores mínimo e máximo de cada classe. Conjuntos de dados sintéticos são construídos de duas maneiras: A primeira forma assume classes com tamanhos diferentes e sem autocorrelação, ver o Algoritmo 1. Uma outra proposta considera a presença de autocorrelação com diferentes valores para o coeficiente de correlação ρ no processo de simulação dos dados.

Considerando o mapa de Munique, dados sintéticos clássicos são gerados para três cenários distintos. Em tais cenários, consideramos baixa, média e alta variabilidade nos dados. No cenário 1, consideramos que a variabilidade dentro de cada região segue uma distribuição uniforme, sendo indicada por $\mathcal{U}(1, 10)$. Para o cenário 2, a variância é dada por uma distribuição $\mathcal{U}(1, 50)$, na qual os dados apresentam uma média concentração de pontos. E por fim, no cenário 3, utilizamos uma distribuição $\mathcal{U}(1, 100)$ para expressar a variabilidade existente. O número de observações geradas para dentro de cada área segue uma distribuição $\mathcal{U}(1, 50)$.

Algoritmo 1: Dados sintéticos intervalares para o mapa da cidade de Munique.

begin
- **Given** um mapa irregular para a cidade de Munique de tamanho n $= 411$;
- **Generate** é gerado um conjunto de dados contínuos $\{q_1, \cdots, q_n\}$, autocorrelacionados ou não autocorrelacionados seguindo uma distribuição normal com média 100 e variância igual a 36. Cada valor contínuo está associado a cada uma das n regiões;
- **for** $k = 1 : n$ **do**
  - **If** o cenário 1 é usado geramos um subconjunto de valores contínuos para a região R_k de acordo com uma distribuição normal com média q_k e variância igual a valores gerados de uma distribuição uniforme no intervalo $[1, 10]$ (**baixa concentração de pontos**). Obtenha o valor mínimo e máximo para a região R_k;
  - **If** o cenário 2 é usado geramos um subconjunto de valores contínuos para a região R_k de acordo com uma distribuição normal com média q_k e variância igual a valores gerados de uma distribuição uniforme no intervalo $[1, 50]$ (**média concentração de pontos**). Obtenha o valor mínimo e máximo para a região R_k;
  - **If** o cenário 3 é usado geramos um subconjunto de valores contínuos para a região R_k de acordo com uma distribuição normal com média q_k e variância igual a valores gerados de uma distribuição uniforme no intervalo $[1, 100]$ (**alta concentração de pontos**). Obtenha o valor mínimo e máximo para a região R_k;
- **end**
- **for** $k = 1 : n$ **do**
  - **for** $t = 1 : n$ **do**
    - **Compute** a matriz de pesos w_{kt} de acordo com o critério descrito na Seção 2;
    - **Compute** a matriz de pesos λ_{kt} de acordo com o critério descrito na Seção 3;
  - **end**
- **end**
- **Compute** o índice de Moran para dados clássicos com a Equação (2.37) e, os $IASi$ e $IASi^*$ com as Equações (3.1) e (3.3), respectivamente.

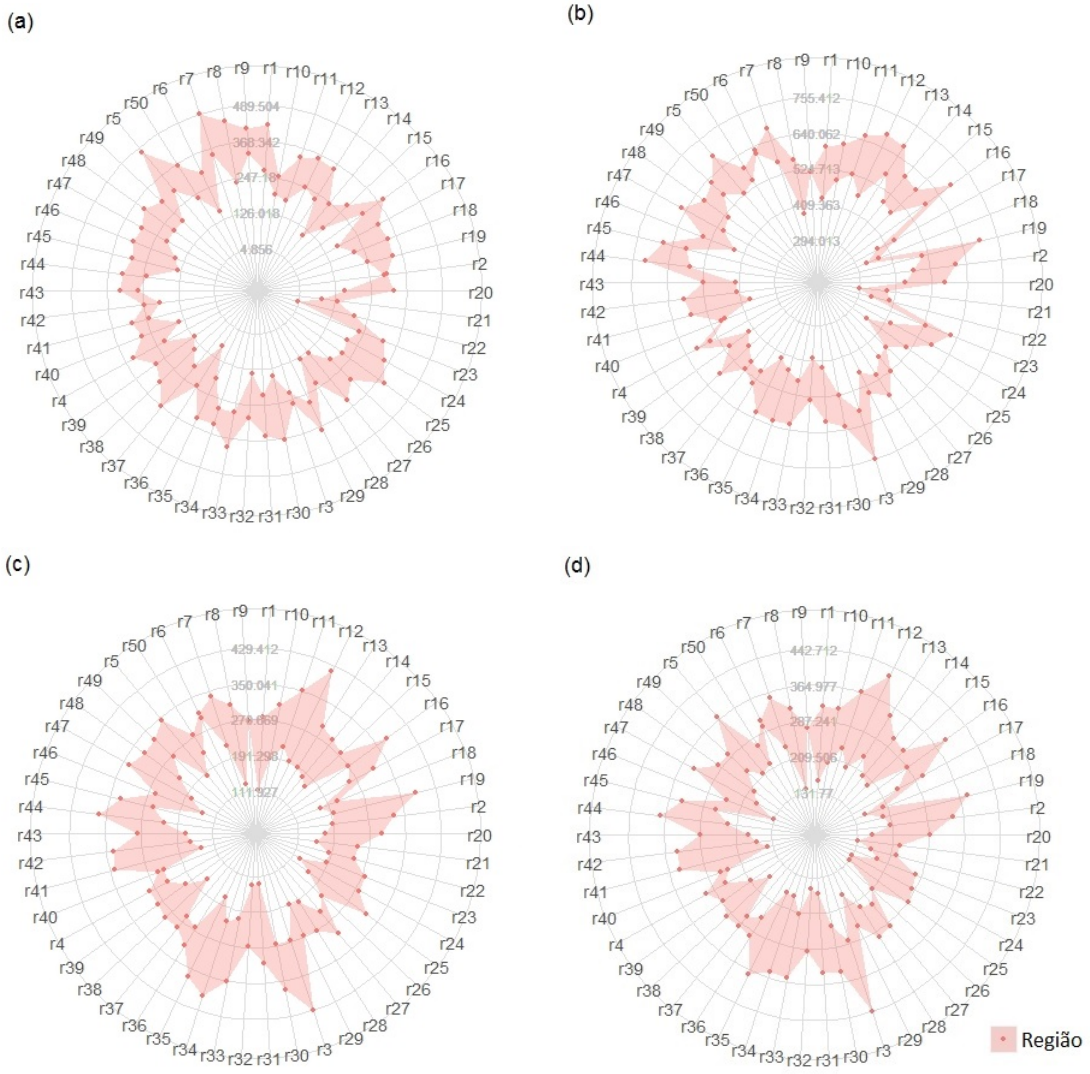
end

A ideia é investigar a variabilidade presente dentro das regiões e analisar o comportamento do índice de autocorrelação espacial para dados intervalares em relação a diferentes representações e duas abordagens para o cálculo das matrizes de pesos. O cálculo da matriz de pesos é uma etapa importante para a obtenção do índice de autocorrelação espacial. Aqui, utilizamos matrizes com diferentes graus de conectividade para verificar a sensibilidade dos resultados em relação ao uso de dados clássicos e intervalares, respectivamente. Em seguida, os experimentos são organizados com base em diferentes matrizes de pesos espaciais.

Dessa maneira, duas configurações de dados intervalares são consideradas neste estudo: dados não autocorrelacionados e dados autocorrelacionados. Esses dados são obtidos a partir de sementes que são valores contínuos não autocorrelacionados ou autocorrelacionados, seguindo a abordagem descrita em Bivand et al. (2008).

De acordo com as simulações geradas com o Algoritmo 1, 144 conjuntos de dados simbólicos foram obtidos. A nível de visualização dos dados simbólicos do tipo intervalo, utilizamos o gráfico de radar para apresentarmos alguns conjuntos de dados. De acordo com a Figura 9, podemos observar a área entre os polígonos superiores e inferiores. Assim, cada eixo representa os limites inferior e superior do intervalo de regiões analisadas, no qual a variabilidade existente dentro de cada classe pode ser observada.

Figura 9 – Dados simulados com diferentes critérios para a geração da matriz de vizinhança: 1-*nn* (a); 2-*nn* (b); distância Chebyshev (c) e contiguidade da rainha (d).



Fonte: Elaborada pela autora.

3.2.1 Resultados para os dados espaciais não-autocorrelacionados

Para uma comparação estatística entre os resultados, são gerados intervalos de confiança para os índices de autocorrelação espacial com o algoritmo de bootstrap (Algoritmo 2). Dessa forma, geramos $B = 1000$ réplicas bootstrap para obtenção dos intervalos bootstrap com tamanhos diferentes. Os intervalos de confiança foram construídos utilizando um nível de confiança igual a 95%. Temos como objetivo analisar os cenários sob diferentes variabilidades. A ideia é criar cenários com baixa concentração de pontos, no qual os dados estão mais dispersos e também cenários com alta concentração de pontos, no qual os dados estão mais concentrados.

Algoritmo 2: Construção dos intervalos de confiança utilizando bootstrap.

begin
  Given Um conjunto de dados reais $\Psi = \{q_1, \ldots, q_n\}$, gerado a partir de uma distribuição normal; O número de amostras bootstrap: B; O valor de significância: α com $0 \leq \alpha \leq 1$;
  Return Intervalo de confiança: $[\underline{c}; \overline{c}]$;
  for *b=1:B* **do**
    amostra$_{boot}$ $\leftarrow$ uma amostra de Ψ, com reposição e tamanho h;
    ind$_{boot}$ $\leftarrow$ índice da amostra$_{boot}$;
  end
  Ordene as índices $ind_{boot1}, ind_{boot2}, \cdots, ind_{bootB}$ em ordem crescente;
  z $\leftarrow$ round $\left(B \cdot \frac{\alpha}{2}\right)$;
  w $\leftarrow$ round $\left[B \cdot \left(1 - \frac{\alpha}{2}\right)\right]$;
  $\underline{c} \leftarrow ind_z$;
  $\overline{c} \leftarrow ind_w$;
end

A Tabela 6 apresenta os intervalos de confiança para os índices de autocorrelação espacial, com 95% de confiança, para o I, IASi e IASi* no contexto de dados simulados. Esses dados são simulados com ausência de autocorrelação espacial, no qual são utilizados quatro critérios de vizinhança. Os valores encontrados destacam que ambos os índices tiveram comportamento semelhantes na ausência de autocorrelação espacial. Para dados do tipo intervalo, o valor do índice aumenta quando a variabilidade (amplitude) aumenta dentro de cada região, exceto para o critério baseado na distância de Chebyshev. Entretanto, isso não é observado para no índice de dados clássicos, pois esse índice não contém informação sobre a variabilidade dentro de cada região. Com relação às representações do centro, centro-amplitude, mínimo-máximo para dados de intervalo, podemos verificar que os valores do índice para a representação centro-amplitude são maiores do que aqueles para o centro e mínimo-máximo quando a variabilidade aumenta. Como esperado, a representação do centro possui os valores mais baixos para o índice de autocorrelação espacial, pois a amplitude é nula nessa representação.

Tabela 6 – Intervalos de confiança com 5% de erro para os índices de autocorrelação espacial, utilizando a matriz de pesos **W**.

Critério de vizinhança	**Dados Clássicos**	**Dados intervalares**		
	I	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
Cenário 1				
1-centroide mais próximo	$[-0,0071;\ -0,0042]$	$[-0,0142;$ 0,0038]	$[-0,0140;$ 0,0048]	$[-0,0148;$ 0,0035]
2-centroides mais próximos	$[-0,0420;\ -0,0399]$	$[-0,0467;\ -0,0337]$	$[-0,0459;\ -0,0325]$	$[-0,0471;\ -0,0342]$
Distância de Chebyshev	[0,0679; 0,0713]	[0,0546; 0,0842]	[0,0514; 0,0825]	[0,0553; 0,0852]
Contiguidade da rainha	$[-0,0457;\ -0,0444]$	$[-0,0462;\ -0,0387]$	$[-0,0455;\ -0,0377]$	$[-0,0464;\ -0,0389]$
Cenário 2				
1-centroide mais próximo	$[-0,0136;$ 0,0005]	$[-0,0375;$ 0,0443]	$[-0,0331;$ 0,0655]	$[-0,0446;$ 0,0375]
2-centroides mais próximos	$[-0,0468;\ -0,0370]$	$[-0,0570;\ -0,0003]$	$[-0,0414;$ 0,0284]	$[-0,0679;-0,0113]$
Distância de Chebyshev	[0,0606; 0,0767]	$[-0,0282;$ 0,1141]	$[-0,0626;$ 0,0808]	$[-0,0223;$ 0,1332]
Contiguidade da rainha	$[-0,0499;\ -0,0437]$	$[-0,0436;\ -0,0108]$	$[-0,0364;$ 0,0041]	$[-0,0497;\ -0,0152]$
Cenário 3				
1-centroide mais próximo	$[-0,0216;$ 0,0060]	$[-0,0487;$ 0,0769]	$[-0,0433;$ 0,1018]	$[-0,0700;$ 0,0709]
2-centroides mais próximos	$[-0,0526;\ -0,0331]$	$[-0,0519;$ 0,0311]	$[-0,0340;$ 0,0695]	$[-0,0811;$ 0,0156
Distância de Chebyshev	[0,0515; 0,0832]	$[-0,1065;$ 0,1052]	$[-0,1332;$ 0,0676]	$[-0,1111;$ 0,1574
Contiguidade da rainha	$[-0,0548;\ -0,0426]$	$[-0,0359;$ 0,0138]	$[-0,0304;$ 0,0303]	$[-0,0479;$ 0,0076

Fonte: Elaborada pela autora.

A Tabela 7 mostra os valores do índice de autocorrelação espacial para dados intervalares, no qual a matriz de pesos utilizada considera a variabilidade existente nas coordenadas geográficas e diferentes critérios para identificar a matriz de vizinhança. A partir desses resultados, podemos observar que os valores estão próximos de zero, resultando na ausência de autocorrelação espacial.

Tabela 7 – Índices de autocorrelação espacial para dados clássicos e intervalar, utilizando a matriz de pesos $\mathbf{\Lambda}$.

Critério de vizinhança	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
	Cenário 1		
1-*nn* com distância de Gowda-Diday	$-0,0340$	$-0,0231$	$-0,0435$
1-*nn* com distância de City block	$-0,0004$	$-0,0086$	0,0067
1-*nn* com distância Euclideana	0,0061	$-0,0037$	0,0148
1-*nn* com distância de Hausdorff	0,0097	$-0,0007$	0,0189
2-*nn* com distância de Gowda-Diday	$-0,0547$	$-0,0502$	$-0,0586$
2-*nn* com distância de City block	$-0,0380$	$-0,0400$	$-0,0361$
2-*nn* com distância Euclideana	$-0,0379$	$-0,0316$	$-0,0434$
2-*nn* com distância de Hausdorff	$-0,0081$	$-0,0183$	0,0007
	Cenário 2		
1-*nn* com distância de Gowda-Diday	0,0346	0,0612	0,0087
1-*nn* com distância de City block	0,0270	0,0157	0,0379
1-*nn* com distância Euclideana	0,0111	0,0032	0,0188
1-*nn* com distância de Hausdorff	$-0,0106$	$-0,0065$	$-0,0146$
2-*nn* com distância de Gowda-Diday	0,0521	0,0515	0,0526
2-*nn* com distância de City block	$-0,0066$	$-0,0171$	0,0034
2-*nn* com distância Euclideana	0,0092	$-0,0104$	0,0284
2-*nn* com distância de Hausdorff	0,0048	$-0,0072$	0,0165
	Cenário 3		
1-*nn* com distância de Gowda-Diday	0,0057	0,0076	0,0039
1-*nn* com distância de City block	$-0,0121$	$-0,0037$	$-0,0198$
1-*nn* com distância Euclideana	0,0008	$-0,0025$	0,0039
1-*nn* com distância de Hausdorff	0,0116	0,0208	0,0032
2-*nn* com distância de Gowda-Diday	0,0254	0,0287	0,0224
2-*nn* com distância de City block	$-0,0120$	$-0,0175$	$-0,0070$
2-*nn* com distância Euclideana	0,0031	0,0001	0,0059
2-*nn* com distância de Hausdorff	$-0,0150$	$-0,0177$	$-0,0126$

Fonte: Elaborada pela autora.

3.2.2 Resultados para os dados espaciais autocorrelacionados

As Figuras 10, 11 e 12 apresentam os resultados do I, IASi, IASi* e IASi*(amplitude = 0) para dados simulados com autocorrelação espacial. Nesses três cenários, todos os índices são calculados usando a matriz de pesos **W**, ou seja, a matriz calculada utilizando distâncias clássicas. A partir dessas figuras, observamos que na presença de autocorrelação espacial, os valores dos índices são superiores aos obtidos com cenário de ausência de autocorrelação espacial. Segundo Bivand et al. (2008) o índice clássico tende a apresentar valores maiores na presença de autocorrelação espacial. Este comportamento também foi observado para os índices no contexto de dados simbólicos intervalares, ou seja, IASi, IASi* e IASi*(amplitude = 0).

Na Figura 10, o valor do índice é calculado considerando a existência de uma baixa variabilidade dentro de cada região, e todos os índices têm comportamentos semelhantes. A

Fig. 11 apresenta uma variabilidade moderada dentro de cada região. Nesse caso, é possível observar comportamentos diferentes entre os índices. Além disso, seus valores para os dados de intervalo são menores do que aqueles apresentados na Figura 10. Em relação aos índices de autocorrelação espacial intervalares, observamos que o IASi* é o mais afetado, pois quando a variabilidade dentro de cada região aumenta, o valor do índice resulta em diferentes situações. Por exemplo, se assumirmos um valor de $\rho = 0,5$ e uma pequena variabilidade dentro de cada região (Fig. 10 (a)), o IASi* obtido é 0,71, indicando uma forte autocorrelação espacial. Diferentemente, quando a variabilidade dentro da região aumenta, obtemos 0,2978 para IASi* (Fig. 11 (a)), indicando uma autocorrelação espacial fraca.

A utilização do índice clássico no cenário de variabilidade moderada dentro de cada região indica uma forte autocorrelação espacial, porém, se usarmos o IASi*, a autocorrelação espacial é moderada. Uma possível justificativa é que o índice para dados de intervalos leva em consideração a variabilidade dentro de cada região. Em geral, o I e o IASi*(amplitude $= 0$) apresentam o mesmo comportamento independente do critério de vizinhança utilizado. Porém, quando o critério de vizinhança utilizado é o 2-centroides mais próximos, os valores são menos impactados, conseguindo capturar a autocorrelação espacial existente. De acordo com a Fig. 12(a), podemos observar um indício de autocorrelação espacial para o IASi*, uma vez que os valores calculados estão próximos de zero. Quando a distância de Chebyshev ou o critério de contiguidade da rainha é utilizado para definição da matriz de vizinhança, independente do valor de ρ utilizado, observamos que os índices IASi e IASi* indicam ausência de autocorrelação espacial.

De acordo com os resultados apresentados nas Figuras 10, 11 e 12, observamos que quando a variabilidade aumenta, os valores dos índices sob diferentes configurações são impactados e as mudanças dependem do tipo de vizinhança e índice utilizado. Nos cenários observados, verificamos que o critério de vizinhança menos afetado é o 2-centroides mais próximo, com a representação IASi*(amplitude $= 0$), que apresentou resultados satisfatórios independentemente da concentração de pontos na região. Com o aumento da variabilidade dentro de cada região, os critérios de vizinhança utilizando a distância de Chebyshev e a contiguidade da rainha foram menos sensíveis, uma vez que os resultados obtidos para os índices indicaram fraca autocorrelação espacial e até mesmo ausência de autocorrelação para as representações consideradas.

Vamos considerar o cenário em que simulamos dados autocorrelacionados, levando em conta a matriz de pesos **W** para o cálculo dos índices. A Tabela 8 apresenta os intervalos

de confiança dos índices de autocorrelação espacial para diferentes valores de ρ, com 95% de confiança. Nesta configuração, os dados apresentam pouca variabilidade entre as regiões. Nessa tabela, os melhores resultados foram obtidos quando utilizamos a matriz de vizinhança 1-*nn*. Nesse cenário, os índices de autocorrelação espacial intervalar conseguiram captar a autocorrelação espacial existente. Por outro lado, os índices de autocorrelação espacial intervalar não apresentaram bons resultados quando utilizamos o critério da rainha para identificar a região de vizinhança.

Quando consideramos uma mudança na variabilidade dos dados (Tabela 9), observamos uma diminuição nos valores dos índices. Esse comportamento é esperado, pois, áreas que apresentam uma maior variabilidade tendem a apresentar uma menor semelhança. A medida que a variabilidade nos dados aumenta (Tabela 10), os índices de autocorrelação espacial intervalar resultam em resultados ruins para alguns critérios de vizinhança. Como exemplo, considere os resultados obtidos com o critério de contiguidade da rainha. Nesse caso, os índices não foram capazes de identificar o tipo de autocorrelação espacial existente nos dados.

Tabela 8 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com baixa concentração de pontos, utilizando a matriz **W**.

Critério de vizinhança	ρ	**Intervalos de confiança (5% de erro)**		
		IASi	**IASi***	**IASi**$^*(amplitude = 0)$
	$-0,6$	[−0,8658; −0,8490]	[−0,8494; −0,8323]	[−0,8717; −0,8546]
	$-0,5$	[−0,7401; −0,7191]	[−0,7199; −0,6987]	[−0,7468; −0,7258]
	$-0,4$	[−0,5988; −0,5749]	[−0,5786; −0,5542]	[−0,6060; −0,5821]
	$-0,3$	[−0,4568; −0,4313]	[−0,4422; −0,4161]	[−0,4624; −0,4361]
	$-0,2$	[−0,3110; −0,2850]	[−0,3007; −0,2741]	[−0,3155; −0,2883]
1-*nn*	$-0,1$	[−0,1674; −0,1406]	[−0,1623; −0,1358]	[−0,1691; −0,1423]
	0,1	[0,1172; 0,14286]	[0,1093; 0,1346]	[0,1196; 0,1456]
	0,2	[0,2539; 0,2982]	[0,2403; 0,2833]	[0,2589; 0,3036]
	0,3	[0,3902; 0,4449]	[0,3722; 0,4252]	[0,3961; 0,4518]
	0,4	[0,5221; 0,5925]	[0,5043; 0,5709]	[0,5289; 0,6001]
	0,5	[0,6445; 0,7332]	[0,6263; 0,7130]	[0,6504; 0,7406]
	0,6	[0,7509; 0,8586]	[0,7368; 0,8416]	[0,7555; 0,8647]
	$-0,6$	[−0,8921; −0,8811]	[−0,8912; −0,8815]	[−0,8937; −0,8859]
	$-0,5$	[−0,8849; −0,8832]	[−0,8824; −0,8806]	[−0,8857; −0,8841]
	$-0,4$	[−0,6313; −0,6238]	[−0,6199; −0,6120]	[−0,6352; −0,6277]
	$-0,3$	[−0,4664; −0,4545]	[−0,4529; −0,4406]	[−0,4710; −0,4594]
	$-0,2$	[−0,3296; −0,3141]	[−0,3181; −0,3023]	[−0,3338; −0,3181]
2-*nn*	$-0,1$	[−0,1977; −0,1785]	[−0,1910; −0,1716]	[−0,2004; −0,1808]
	0,1	[0,0869; 0,11909]	[0,0807; 0,1120]	[0,0891; 0,1218]
	0,2	[0,2447; 0,29681]	[0,2322; 0,2821]	[0,2484; 0,3015]
	0,3	[0,3998; 0,50117]	[0,3886; 0,4851]	[0,4037; 0,5069]
	0,4	[0,5252; 0,7330]	[0,5193; 0,7231]	[0,5272; 0,7364]
	0,5	[0,5423; 0,8814]	[0,5420; 0,8807]	[0,5424; 0,8816]
	0,6	[0,5728; 0,8987]	[0,5711; 0,8965]	[0,5711; 0,8889]
	$-0,6$	[−0,3084; −0,2637]	[−0,2955; −0,2506]	[−0,3136; −0,2678]
	$-0,5$	[−0,2013; −0,1621]	[−0,1928; −0,1536]	[−0,2042; −0,1650]
	$-0,4$	[−0,2131; −0,1529]	[−0,1421; −0,1265]	[−0,1798; −0,1330]
	$-0,3$	[−0,0690; −0,0352]	[−0,0689; −0,0337]	[−0,0700; −0,0355]
	$-0,2$	[−0,0239; 0,0107]	[−0,0260; 0,0101]	[−0,0233; 0,0113]
Distância de	$-0,1$	[0,0151; 0,0518]	[0,0114; 0,0494]	[0,0161; 0,0530]
Chebyshev	0,1	[0,0911; 0,1340]	[0,0837; 0,1272]	[0,0937; 0,1367]
	0,2	[0,1329; 0,1826]	[0,1235; 0,1735]	[0,1356; 0,1861]
	0,3	[0,1797; 0,2402]	[0,1690; 0,2287]	[0,1835; 0,2446]
	0,4	[0,2352; 0,3113]	[0,2231; 0,2965]	[0,2399; 0,3163]
	0,5	[0,3007; 0,4023]	[0,2871; 0,3839]	[0,3053; 0,4088]
	0,6	[0,3757; 0,5276]	[0,3595; 0,5049]	[0,3809; 0,5353]
	$-0,6$	[−0,1862; −0,1782]	[−0,1802; −0,1717]	[−0,1883; −0,1803]
	$-0,5$	[−0,1681; −0,1593]	[−0,1628; −0,1535]	[−0,1700; −0,1612]
	$-0,4$	[−0,1492; −0,1394]	[−0,1446; −0,1346]	[−0,1509; −0,1411]
	$-0,3$	[−0,1292; −0,1184]	[−0,1255; −0,1146]	[−0,1308; −0,1197]
	$-0,2$	[−0,1080; −0,0957]	[−0,1052; −0,0929]	[−0,1091; −0,0966]
Contiguidade	$-0,1$	[−0,0854; −0,0713]	[−0,0835; −0,0699]	[−0,0863; −0,0719]
da rainha	0,1	[−0,0349; −0,0160]	[−0,0356; −0,0172]	[−0,0348; −0,0156]
	0,2	[−0,0066; 0,0156]	[−0,0086; 0,0130]	[−0,0060; 0,0168]
	0,3	[0,0241; 0,0516]	[0,0205; 0,0472]	[0,0249; 0,0533]
	0,4	[0,0567; 0,0922]	[0,0523; 0,0860]	[0,0583; 0,0945]
	0,5	[0,0918; 0,1401]	[0,0864; 0,1318]	[0,0937; 0,1429]
	0,6	[0,1268; 0,1956]	[0,1209; 0,1859]	[0,1288; 0,1990]

Fonte: Elaborada pela autora.

Tabela 9 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com média concentração de pontos, utilizando a matriz **W**.

Critério de vizinhança	ρ	**Intervalos de confiança (5% de erro)**		
		IASi	**IASi***	**IASi**$^*(amplitude = 0)$
	$-0,6$	[−0,7046; −0,6554]	[−0,5264; −0,4535]	[−0,8079; −0,7610]
	$-0,5$	[−0,5636; −0,5005]	[−0,3991; −0,3135]	[−0,6758; −0,6126]
	$-0,4$	[−0,4363; −0,3621]	[−0,3023; −0,2084]	[−0,5410; −0,4625]
	$-0,3$	[−0,3643; −0,2881]	[−0,2649; −0,1644]	[−0,4441; −0,3609]
	$-0,2$	[−0,2657; −0,1834]	[−0,2012; −0,0991]	[−0,3240; −0,2353]
1-*nn*	$-0,1$	[−0,1756; −0,0922]	[−0,1456; −0,0430]	[−0,2094; −0,1182]
	0,1	[0,0390; 0,1255]	[−0,0231; 0,0815]	[0,0740; 0,1806]
	0,2	[0,1308; 0,2177]	[0,0308; 0,1346]	[0,1952; 0,2974]
	0,3	[0,2315; 0,3164]	[0,0942; 0,1967]	[0,3227; 0,4206]
	0,4	[0,3474; 0,4286]	[0,1766; 0,2734]	[0,4529; 0,5469]
	0,5	[0,4753; 0,5578]	[0,2839; 0,3744]	[0,5837; 0,6782]
	0,6	[0,6110; 0,6992]	[0,4222; 0,5064]	[0,7026; 0,8059]
	$-0,6$	[−0,8677; −0,8522]	[−0,8198; −0,7901]	[−0,8997; −0,8592]
	$-0,5$	[−0,8554; −0,8457]	[−0,8022; −0,7855]	[−0,8756; −0,8667]
	$-0,4$	[−0,5220; −0,4870]	[−0,3959; −0,3449]	[−0,5946; −0,5604]
	$-0,3$	[−0,3569; −0,3116]	[−0,2559; −0,1950]	[−0,4309; −0,3825]
	$-0,2$	[−0,2499; −0,1990]	[−0,1813; −0,1160]	[−0,3097; −0,2517]
2-*nn*	$-0,1$	[−0,1641; −0,1082]	[−0,1295; −0,0617]	[−0,2001; −0,1357]
	0,1	[0,0104; 0,0723]	[−0,0334; 0,0373]	[0,0352; 0,1102]
	0,2	[0,1266; 0,1907]	[0,0382; 0,1069]	[0,1810; 0,2594]
	0,3	[0,2781; 0,3601]	[0,1548; 0,2269]	[0,3491; 0,4509]
	0,4	[0,4607; 0,6281]	[0,3653; 0,4858]	[0,5027; 0,6973]
	0,5	[0,5379; 0,8753]	[0,5297; 0,8602]	[0,5407; 0,8801]
	0,6	[0,5498; 0,8895]	[0,5337; 0,8733]	[0,5575; 0,8973]
	$-0,6$	[−0,2485; −0,1287]	[−0,1907; −0,0621]	[−0,3168; −0,1757]
	$-0,5$	[−0,1818; −0,0695]	[−0,1586; −0,0280]	[−0,2253 −0,0954]
	$-0,4$	[−0,0712; 0,0252]	[−0,0662; 0,0181]	[−0,1070; 0,0357]
	$-0,3$	[−0109; 0,0100]	[−0,1175; 0,0143]	[−0,1238; 0,0166]
	$-0,2$	[−0,0855; 0,0395]	[−0,1046; 0,0309]	[−0,0901; 0,0573]
Distância de	$-0,1$	[−0,0650 0,0668]	[−0,0930; 0,0459]	[−0,0603 0,0966]
Chebyshev	0,1	[−0,0247; 0,1246]	[−0,0699; 0,0757]	[−0,0079; 0,1773]
	0,2	[−0,0001; 0,158]	[−0,0559; 0,0930]	[0,0238; 0,2235]
	0,3	[0,0287; 0,2003]	[−0,0371; 0,1163]	[0,0585; 0,2761]
	0,4	[0,0638; 0,2528]	[−0,0163; 0,1481]	[0,1060; 0,3457]
	0,5	[0,1120; 0,3212]	[0,0156; 0,1886]	[0,1656; 0,4311]
	0,6	[0,1789; 0,4111]	[0,0618; 0,2462]	[0,2464; 0,5395]
	$-0,6$	[−0,1477; −0,1160]	[−0,1128; −0,0752]	[−0,1786; -0,1430]
	$-0,5$	[−0,1350; −0,1025]	[−0,1049; −0,0671]	[−0,1637; −0,1256]
	$-0,4$	[−0,1225; −0,0890]	[−0,0975; −0,0590]	[−0,1482; −0,1079]
	$-0,3$	[−0,1102; −0,0754]	[−0,0900; −0,0511]	[−0,1317; −0,0895]
	$-0,2$	[−0,0972; −0,0615]	[−0,0825; −0,0434]	[−0,1143; −0,0704]
Contiguidade	$-0,1$	[−0,0837; −0,0472]	[−0,0750; −0,03543]	[−0,0958; −0,0506]
da rainha	0,1	[−0,0536; −0,0140]	[−0,0580; −0,01800]	[−0,0534; −0,0062]
	0,2	[−0,0359; 0,0045]	[−0,0481; −0,0078]	[−0,02923; 0,0192]
	0,3	[−0,0156; 0,0260]	[−0,0365; 0,0044]	[−0,00235; 0,0495]
	0,4	[0,0077; 0,0514]	[−0,0232; 0,0192]	[0,02777; 0,0844]
	0,5	[0,0356; 0,0833]	[−0,0056; 0,0379]	[0,06232; 0,1230]
	0,6	[0,0680; 0,1237]	[0,0175; 0,0626]	[0,0988; 0,1720]

Fonte: Elaborada pela autora.

Tabela 10 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com alta concentração de pontos, utilizando a matriz **W**.

Critério de vizinhança	ρ	**Intervalos de confiança (5% de erro)**		
		IASi	**IASi***	**IASi**$^*(amplitude = 0)$
	$-0,6$	[−0,4591; −0,3706]	[−0,2794; −0,1675]	[−0,6581; −0,5628]
	$-0,5$	[−0,3416; −0,2414]	[−0,2117; −0,0928]	[−0,5196; −0,4031]
	$-0,4$	[−0,2560; −0,1517]	[−0,1698; −0,0473]	[−0,4022; −0,2673]
	$-0,3$	[−0,2390; −0,1289]	[−0,1678; −0,0396]	[−0,3536; −0,2195]
	$-0,2$	[−0,1876; −0,0780]	[−0,1433; −0,0150]	[−0,2725; −0,1358]
	$-0,1$	[−0,1443; −0,0334]	[−0,1241; 0,0049]	[−0,2006; −0,0625]
1-*nn*	0,1	[−0,0267; 0,0813]	[−0,0737; 0,0518]	[0,0128; 0,1632]
	0,2	[−0,0554; 0,0697]	[0,0903; 0,2413]	[0,1292; 0,2886]
	0,3	[0,0688; 0,1738]	[0,0318; 0,0920]	[0,1792; 0,3261]
	0,4	[0,1371; 0,2415]	[0,0013; 0,1243]	[0,2841; 0,4222]
	0,5	[0,2309; 0,3311]	[0,0502; 0,1726]	[0,4135; 0,5389]
	0,6	[0,3583; 0,4534]	[0,1330; 0,2493]	[0,5549; 0,6705]
	$-0,6$	[−0,8122; −0,7932]	[−0,7344; −0,6945]	[−0,8821; −0,8322]
	$-0,5$	[−0,7735; −0,7489]	[−0,6291; −0,5863]	[−0,8427; −0,8216]
	$-0,4$	[−0,3492; −0,2864]	[−0,2196; −0,1428]	[−0,4891; −0,4214]
	$-0,3$	[−0,2244; −0,1560]	[−0,1512; −0,0688]	[−0,3338; −0,2511]
	$-0,2$	[−0,1633; −0,0919]	[−0,1228; −0,0379]	[−0,2399; −0,1504]
2-*nn*	$-0,1$	[−0,1217; −0,0485]	[−0,1058; −0,0197]	[−0,1694; −0,0712]
	0,1	[−0,0466; 0,02825]	[−0,0743; 0,0124]	[−0,0270; 0,0787]
	0,2	[0,0094; 0,0856]	[−0,0502; 0,0353]	[0,0734; 0,1814]
	0,3	[0,1118; 0,1904]	[0,0003; 0,0835]	[0,2236; 0,3345]
	0,4	[0,3231; 0,4349]	[0,1561; 0,2416]	[0,4353; 0,5975]
	0,5	[0,5247; 0,8529]	[0,4989; 0,7939]	[0,5356; 0,8726]
	0,6	[0,5421; 0,8795]	[0,5243; 0,8122]	[0,5399; 0,8792]
	$-0,6$	[−0,1765; −0,0300]	[−0,1501; 0,0059]	[−0,2723; −0,0630]
	$-0,5$	[−0,1498; −0,0007]	[−0,1388; 0,0159]	[−0,2229; −0,0045]
	$-0,4$	[−0,0570; 0,1617]	[−0,1006; 0,0770]	[−0,0535; 0,3066]
	$-0,3$	[−0,1232; 0,0366]	[−0,1261; 0,0324]	[−0,1660; 0,0706]
	$-0,2$	[−0,1142; 0,0515]	[−0,1230; 0,0370]	[−0,1507; 0,0989]
Distância de	$-0,1$	[−0,1064; 0,0651]	[−0,1206; 0,0423]	[−0,1357; 0,1241]
Chebyshev	0,1	[−0,0907; 0,0943]	[−0,1159; 0,0516]	[−0,1134; 0,1815]
	0,2	[−0,0834; 0,1111]	[−0,1127; 0,0574]	[−0,0983; 0,2169]
	0,3	[−0,0741; 0,1332]	[−0,1077; 0,0652]	[−0,0808; 0,2564]
	0,4	[−0,0570; 0,1617]	[−0,1006; 0,0770]	[−0,0535; 0,3066]
	0,5	[−0,0365; 0,1991]	[−0,0898; 0,0916]	[−0,0165; 0,3691]
	0,6	[0,0003; 0,2514]	[−0,0730; 0,1141]	[0,0423; 0,4542]
	$-0,6$	[−0,1028; −0,0604]	[−0,0820; −0,0348]	[−0,1438; −0,0835]
	$-0,5$	[−0,0964; −0,0536]	[−0,0794; −0,0319]	[−0,1335; −0,0718]
	$-0,4$	[−0,0904; −0,0469]	[−0,0769; −0,0292]	[−0,1233; −0,0603]
	$-0,3$	[−0,0846; −0,0408]	[−0,0746; −0,0266]	[−0,1133; −0,0481]
	$-0,2$	[−0,0791; −0,0348]	[−0,0723; −0,0241]	[−0,1031; −0,0366]
Contiguidade	$-0,1$	[−0,0734; −0,0287]	[−0,0700; −0,0216]	[−0,0913; −0,0235]
da rainha	0,1	[−0,0608; −0,0147]	[−0,0647; −0,0159]	[−0,0672; 0,0037]
	0,2	[−0,0535; −0,0061]	[−0,0618; −0,0126]	[−0,0529; 0,0201]
	0,3	[−0,0446; 0,0031]	[−0,0577; −0,0088]	[−0,0360; 0,0390]
	0,4	[−0,0337; 0,0150]	[−0,0530; −0,0038]	[−0,0172; 0,0600]
	0,5	[−0,0200; 0,0301]	[−0,0470; 0,0018]	[0,0072; 0,0866]
	0,6	[−0,0017; 0,0513]	[−0,0381; 0,0111]	[0,0376; 0,1226]

Fonte: Elaborada pela autora.

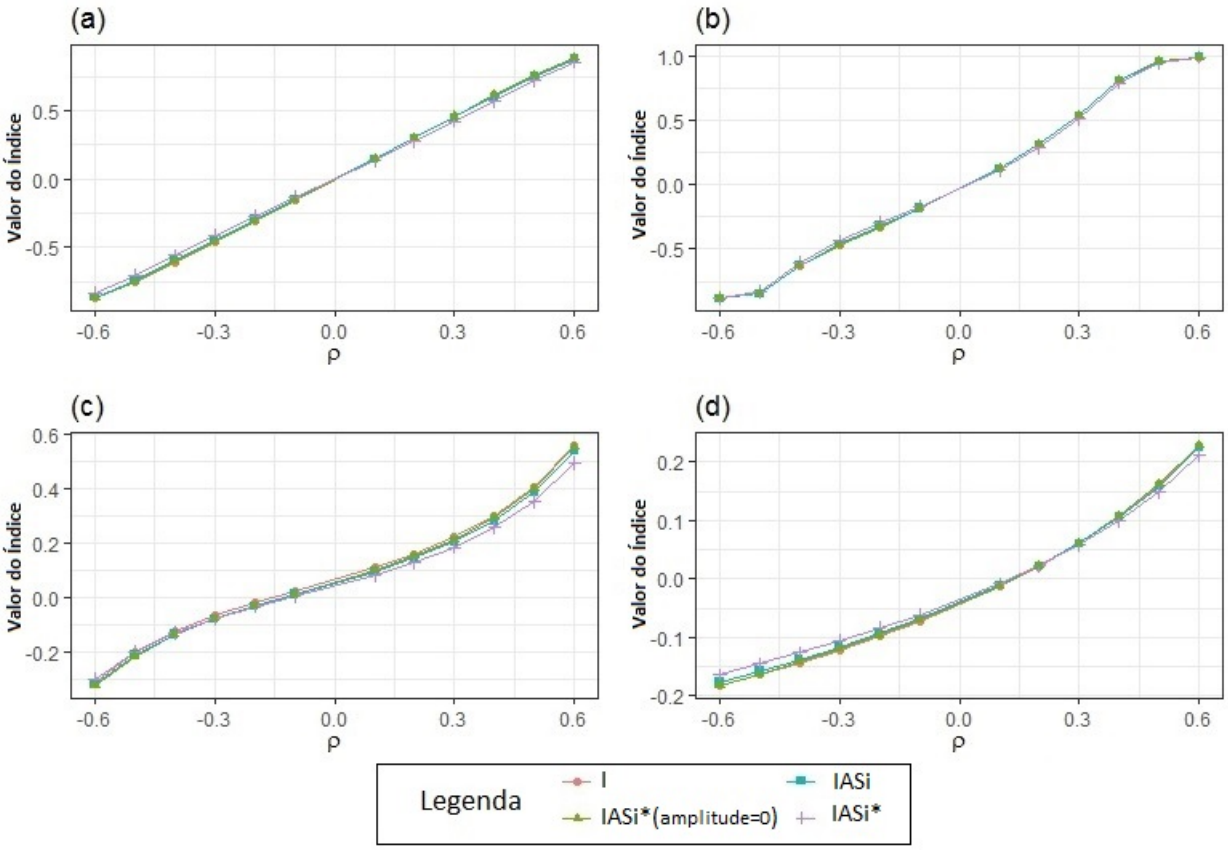
Na Figura 13, podemos ver que o aumento no número de vizinhos interfere no aumento

dos índices para dados de intervalos (Fig. 13). Como esperado, o IASi* (amplitude$=0$) apresenta os maiores valores, quando comparado com as demais representações. Comparamos os índices relativos referentes às matrizes **W** (Figs. 10(a) e 10(b)) com os índices relacionados às matrizes $\mathbf{\Lambda}$. Podemos ver que os valores dos índices baseados na matriz $\mathbf{\Lambda}$ são menores. Esse comportamento está relacionado ao fato de estarmos lidando com áreas com tamanhos pequenos e formas irregulares. Então, como esperado, a variabilidade existente entre as coordenadas geográficas não implica valores maiores para o índices.

Para analisarmos o comportamento da autocorrelação espacial nos dados, usamos a ideia de Bootstrap. Nesse cenário, simulamos dados autocorrelacionados quando consideramos a matriz de pesos $\mathbf{\Lambda}$. Aqui, a matriz de pesos é construída utilizando distâncias intervalares. Dessa maneira, a variabilidade existente entre as coordenadas geográficas também é levada em consideração.

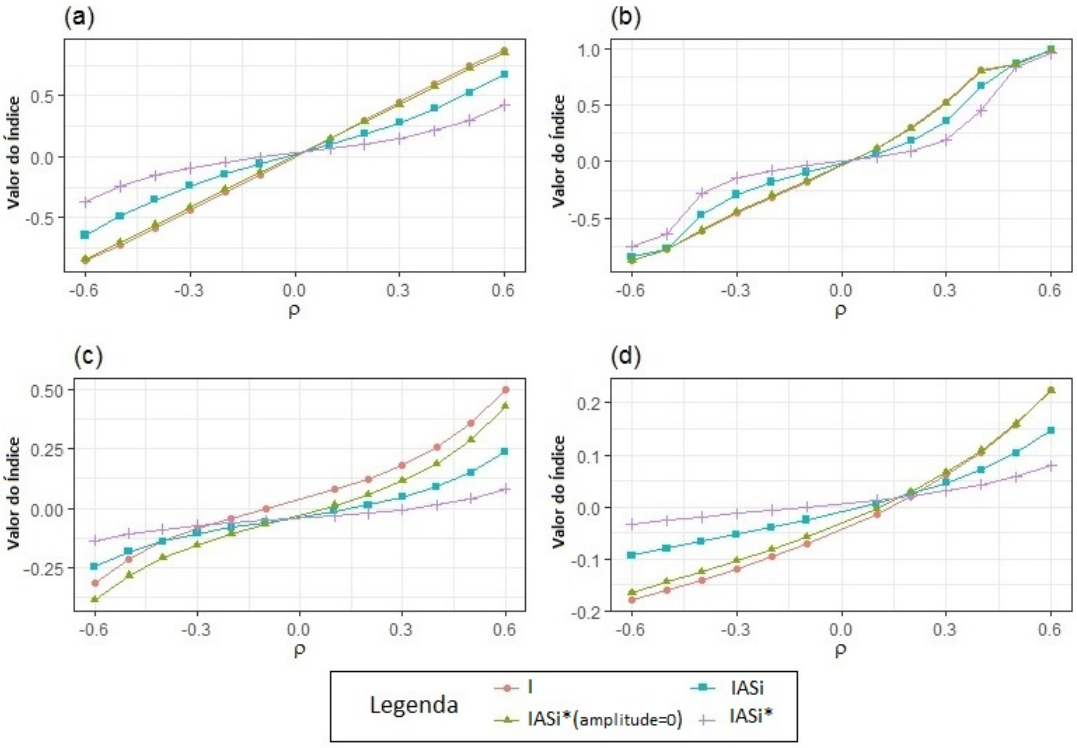
Intervalos de confiança para os índices de autocorrelação espacial sob diferentes variabilidades e valores de ρ foram utilizados. Devido ao custo computacional da simulação, independente do índice utilizado, consideramos os seguintes valores de ρ: -0,3; -0,2; -0,1; 0,1; 0,2 e 0,3. Na Tabela 11, observamos que os valores encontrados são semelhantes. A medida que a variabilidade aumenta, os valores diminuem. Esse comportamento é comum para as demais distâncias intervalares analisadas (Tabelas 12, 13 e 14). Dentre os critérios analisados, os índices de autocorrelação espacial intervalares utilizando a distância de Hausdorff apresentaram os melhores reultados.

Figura 10 – Valores para os índices utilizando o critério de vizinhança: 1-*nc* (a), 2-*nc* (b), distância de Chebyshev (c) e contiguidade da rainha (d), baseado na matriz **W** com variabilidade no geo-objeto baixa.



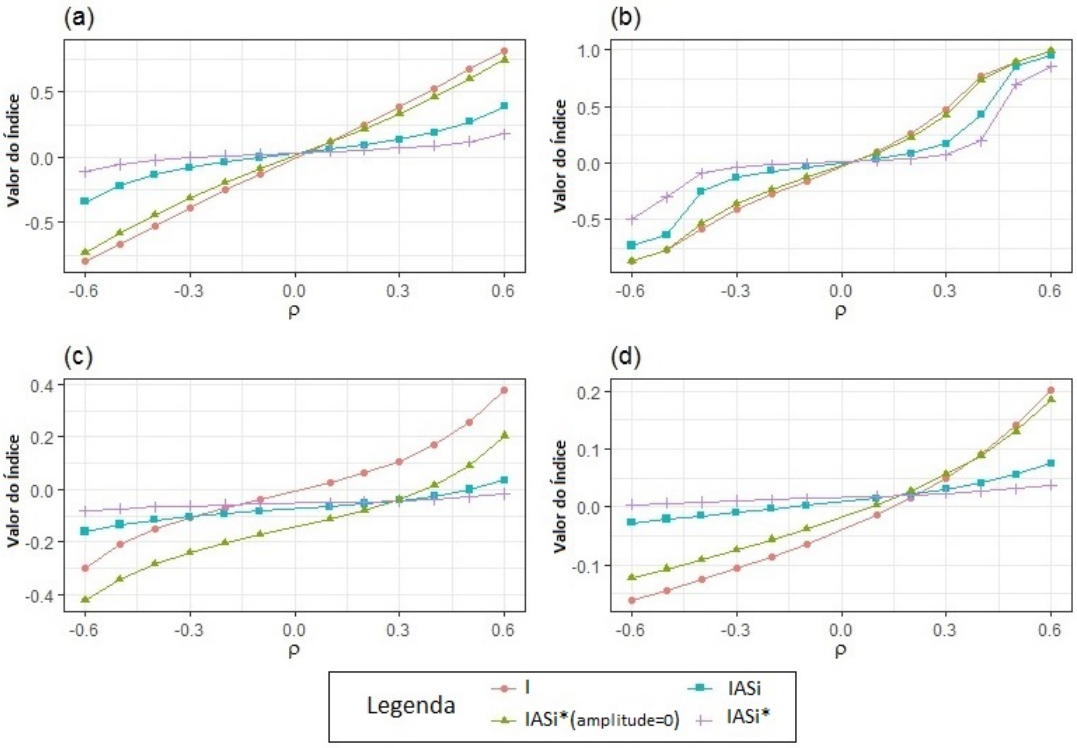
Fonte: Elaborada pela autora.

Figura 11 – Valores para os índices utilizando o critério de vizinhança: 1-*nc* (a), 2-*nc*(b), distância de Chebyshev (c) e contiguidade da rainha (d), baseados na matriz **W** com variabilidade no geo-objeto moderada.



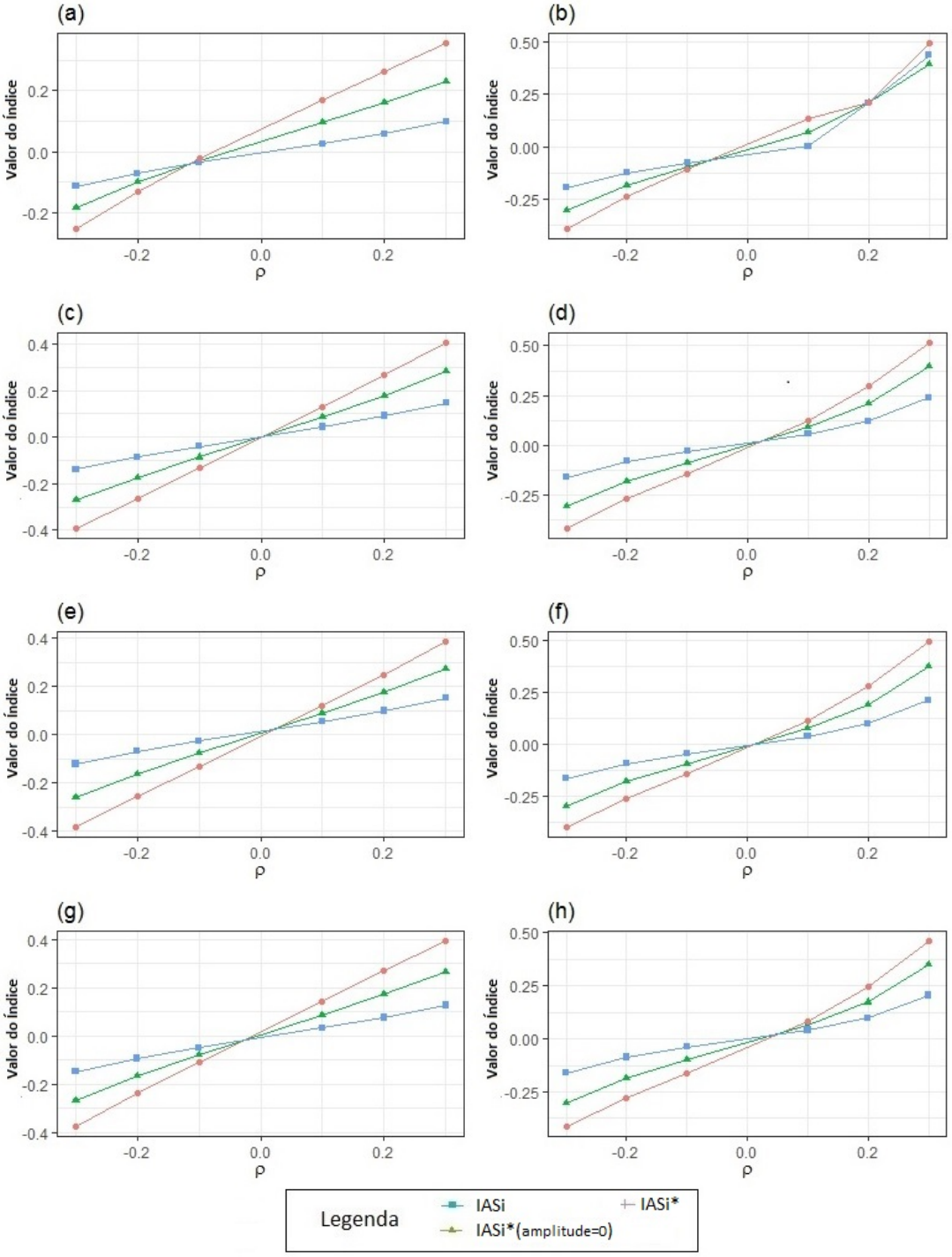
Fonte: Elaborada pela autora.

Figura 12 – Valores para os índices utilizando os critérios de vizinhança: 1-*nc* (a), 2-*nc* (b), distância de Chebyshev (c) e contiguidade da rainha (d), baseado na matriz **W** com variabilidade no geo-objeto alta.



Fonte: Elaborada pela autora.

Figura 13 – Valores para o $IASi$, $IASi^*$ e $IASi^*$(amplitude=0) com a matriz $\mathbf{\Lambda}$, utilizando $\mathbf{1}$-*nn* com as distâncias: Gowda-Diday, City block, Euclideana e Hausdorff em (a), (c), (e) e (g), respectivamente; $\mathbf{2}$-*nc* com as distâncias de Gowda-Diday, City block, Euclideana e Hausdorff em (b), (d), (f) e (h), respectivamente.



Fonte: Elaborada pela autora.

Tabela 11 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com diferentes variabilidades, utilizando o K-*nn* com a distância de Gowda-Diday para o cálculo da matriz $\mathbf{\Lambda}$.

Intervalos de confiança (5% de erro)			
1-***nn*** **com distância de Gowda-Diday**			
ρ	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
$-0,3$	[−0,3880; −0,3637]	[−0,3692; −0,3440]	[−0,3946; −0,3704]
$-0,2$	[−0,2498; −0,2237]	[−0,2379; −0,2129]	[−0,2540;−0,2281]
$-0,1$	[−0,1033; −0,0776]	[−0,0963; −0,0702]	[−0,1061; −0,0799]
0,1	[0,1372; 0,1683]	[0,1331; 0,1630]	[0,1386; 0,1703]
0,2	[0,2463; 0,2831]	[0,2378; 0,2731]	[0,2493; 0,2871]
0,3	[0,3515; 0,3974]	[0,3393; 0,3834]	[0,3559; 0,4026]
$-0,3$	[−0,2666; −0,1924]	[−0,1482; −0,0625]	[−0,3693; −0,2829]
$-0,2$	[−0,1927; −0,1181]	[−0,1297; −0,0304]	[−0,2537; −0,1691]
$-0,1$	[−0,0837; −0,0084]	[−0,0414; 0,0458]	[−0,1293; −0,0378]
$0,1$	[0,0658; 0,1401]	[0,04078; 0,1281]	[0,0753; 0,1638]
$0,2$	[0,1396; 0,2115]	[0,08393; 0,1691]	[0,1725; 0,2597]
$0,3$	[0,2192; 0,2881]	[0,13315; 0,2152]	[0,2714; 0,3583]
$-0,3$	[−0,1358; −0,0391]	[−0,0561; 0,0544]	[−0,2804; −0,1481]
$-0,2$	[−0,0877; 0,0089]	[−0,0364; 0,0757]	[−0,1973; −0,0615]
$-0,1$	[−0,0481; 0,0465]	[−0,0204; 0,0916]	[−0,1264; 0,0135]
$0,1$	[0,0177; 0,1117]	[0,0061; 0,1175]	[0,0014; 0,1372]
$0,2$	[0,0509; 0,1438]	[0,0203; 0,1308]	[0,0668; 0,2016]
$0,3$	[0,0897; 0,1823]	[0,0374; 0,1461]	[0,1384; 0,2672]
2-***nn*** **com distância de Gowda-Diday**			
ρ	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
$-0,3$	[−0,4607; −0,4494]	[−0,4449; −0,4323]	[−0,4664; −0,4552]
$-0,2$	[−0,2984; −0,2815]	[−0,2878; −0,2703]	[−0,3023; −0,2851]
$-0,1$	[−0,1441; −0,1281]	[−0,1366; −0,1193]	[−0,1470; −0,1310]
0,1	[0,1410; 0,1680]	[0,13597; 0,1615]	[0,1430; 0,1702]
0,2	[0,2872; 0,3303]	[0,27762; 0,3180]	[0,2908; 0,3348]
0,3	[0,4462; 0,5307]	[0,43600; 0,5164]	[0,4497; 0,5359]
$-0,3$	[−0,3348; −0,2880]	[−0,2112; −0,1489]	[−0,4252; −0,3747]
$-0,2$	[−0,2167; −0,1651]	[−0,1536; −0,0880]	[−0,2684; −0,2126]
$-0,1$	[−0,1050; −0,0514]	[−0,0634; 0,0018]	[−0,1492; −0,0867]
0,1	[0,0702; 0,1250]	[0,0348; 0,0999]	[0,0891; 0,1548]
0,2	[0,1762; 0,2320]	[0,0993; 0,1639]	[0,2239; 0,2910]
0,3	[0,3324; 0,3975]	[0,2170; 0,2829]	[0,3942; 0,4769]
$-0,3$	[−0,1788; −0,1114]	[−0,0891; −0,0043]	[−0,3188; −0,2319]
$-0,2$	[−0,1050; −0,0366]	[−0,0535; 0,0309]	[−0,2044; −0,1104]
$-0,1$	[−0,0574; 0,0110]	[−0,0330; 0,0509]	[−0,1207; −0,0243]
0,1	[0,0190; 0,0872]	[−0,0019; 0,0808]	[0,0256; 0,1235]
0,2	[0,0705; 0,1404]	[0,0204; 0,1029]	[0,1204; 0,2158]
0,3	[0,1683; 0,2411]	[0,0694; 0,1494]	[0,2691; 0,3642]

Fonte: Elaborada pela autora.

Tabela 12 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com diferentes variabilidades, utilizando o k-*nn* com a distância de City block para o cálculo da matriz $\mathbf{\Lambda}$.

Intervalos de confiança (5% de erro)			
1-***nn*** **com distância de City block**			
ρ	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
$-0,3$	[−0,50479; −0,48004]	[−0,48866; −0,46204]	[−0,51078; −0,48623]
$-0,2$	[−0,35705; −0,32846]	[−0,34454; −0,31620]	[−0,36139; −0,33234]
$-0,1$	[−0,20283; −0,17122]	[−0,19738; −0,16604]	[−0,20512; −0,17285]
0,1	[0,11356; 0,15093]	[0,10314; 0,14016]	[0,11678; 0,15461]
0,2	[0,27044; 0,31183]	[0,25323; 0,29493]	[0,27627; 0,31821]
0,3	[0,42074; 0,46960]	[0,40025; 0,44822]	[0,42739; 0,47733]
$-0,3$	[−0,38319; −0,31293]	[−0,28444; −0,18730]	[−0,47040; −0,39497]
$-0,2$	[−0,27981; −0,20139]	[−0,22037; −0,12113]	[−0,34077; −0,25437]
$-0,1$	[−0,18007; −0,09773]	[−0,16398; −0,06260]	[−0,20774; −0,11373]
0,1	[0,01700; 0,10188]	[−0,05477; 0,04662]	[0,06285; 0,16130]
0,2	[0,12131; 0,20709]	[0,00584; 0,10840]	[0,20030; 0,29841]
0,3	[0,23567; 0,32007]	[0,08035; 0,18061]	[0,33871; 0,43589]
$-0,3$	[−0,24784; −0,14854]	[−0,19155; −0,06733]	[−0,36582; −0,23905]
$-0,2$	[−0,19432; −0,09201]	[−0,16834; −0,04243]	[−0,27360; −0,13587]
$-0,1$	[−0,14772; −0,04453]	[−0,14884; −0,02383]	[−0,18732; −0,04454]
0,1	[−0,06091; 0,04593]	[−0,11200; 0,01390]	[−0,02084; 0,12920]
0,2	[−0,01069; 0,09485]	[−0,09146; 0,03392]	[0,07215; 0,22167]
0,3	[0,23567; 0,32007]	[0,08035; 0,18061]	[0,33871; 0,43589]
2-***nn*** **com distância de City block**			
ρ	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
$-0,3$	[−0,50824; −0,49382]	[−0,49424; −0,47996]	[−0,51334; −0,49854]
$-0,2$	[−0,35192; −0,33218]	[−0,33988; −0,31984]	[−0,35644; −0,33629]
$-0,1$	[−0,20196; −0,17566]	[−0,19527; −0,16943]	[−0,20461; −0,17821]
0,1	[0,12841; 0,16838]	[0,12013; 0,15854]	[0,13127; 0,17223]
0,2	[0,31290; 0,36846]	[0,29945; 0,35240]	[0,31761; 0,37410]
0,3	[0,49612; 0,58724]	[0,48271; 0,56985]	[0,50067; 0,59333]
$-0,3$	[−0,38435; −0,33471]	[−0,27517; −0,20693]	[−0,46076; −0,41327]
$-0,2$	[−0,25468; −0,19579]	[−0,18587; −0,11016]	[−0,31108; −0,25224]
$-0,1$	[−0,15057; −0,08674]	[−0,12392; −0,04400]	[-0,17862; −0,10941]
0,1	[0,05851; 0,12943]	[−0,00602; 0,07589]	[0,10224; 0,18584]
0,2	[0,19675; 0,27107]	[0,08194; 0,16230]	[0,27220; 0,35865]
0,3	[0,37596; 0,45955]	[0,22616; 0,30419]	[0,45872; 0,55739]
$-0,3$	[−0,23496; −0,15718]	[−0,15813; −0,06370]	[−0,34977; −0,26389]
$-0,2$	[−0,15674; −0,07518]	[−0,12122; −0,02344]	[−0,23039; −0,12940]
$-0,1$	[−0,10599; −0,02159]	[−0,09863; −0,00110]	[−0,13941; −0,02981]
0,1	[−0,01235; 0,07625]	[−0,06038; 0,03975]	[0,03818; 0,15970]
0,2	[0,05990; 0,14820]	[−0,02779; 0,07094]	[0,16109; 0,28249]
0,3	[0,37596; 0,45955]	[0,22616; 0,30419]	[0,45872; 0,55739]

Fonte: Elaborada pela autora.

Tabela 13 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com diferentes variabilidades, utilizando o k-*nn* com a distância euclideana para o cálculo da matriz $\mathbf{\Lambda}$.

Intervalos de confiança (5% de erro)			
	1-***nn*** **com distância euclideana**		
ρ	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
$-0,3$	[−0,50944; −0,48804]	[−0,49194; −0,47004]	[−0,51549; −0,49421]
$-0,2$	[−0,36389; −0,33823]	[−0,35143; −0,32557]	[−0,36873; −0,34272]
$-0,1$	[−0,21398; −0,18484]	[−0,20790; −0,17878]	[−0,21661; −0,18735]
0,1	[0,09346; 0,12908]	[0,08466; 0,11968]	[0,09668; 0,13293]
0,2	[0,24682; 0,28731]	[0,23128; 0,27105]	[0,25198; 0,29390]
0,3	[0,39647; 0,44535]	[0,37680; 0,42393]	[0,40335; 0,45314]
$-0,3$	[−0,38769; −0,31572]	[−0,28167; −0,19036]	[−0,47659; −0,40072]
$-0,2$	[−0,27025; −0,18964]	[−0,19884; −0,09869]	[−0,33261; −0,24044]
$-0,1$	[−0,19026; −0,10856]	[−0,16324; −0,06572]	[−0,22028; −0,12922]
0,1	[0,00061; 0,08454]	[−0,05877; 0,04092]	[0,03597; 0,13677]
0,2	[0,10104; 0,18752]	[−0,00037; 0,10136]	[0,17147; 0,27065]
0,3	[0,21327; 0,29883]	[0,06861; 0,16817]	[0,31005; 0,40749]
$-0,3$	[−0,25385; −0,15109]	[−0,18704; −0,06957]	[−0,37542; −0,24357]
$-0,2$	[−0,20065; −0,09594]	[−0,16420; −0,04473]	[−0,28669; −0,14593]
$-0,1$	[−0,15622; −0,04986]	[−0,14483; −0,02525]	[−0,20309; −0,05603]
0,1	[−0,07322; 0,03571]	[−0,11125; 0,00961]	[−0,04302; 0,10772]
0,2	[−0,02691; 0,08297]	[−0,09180; 0,02976]	[0,04433; 0,19526]
0,3	[0,03057; 0,13963]	[−0,06716; 0,05479]	[0,14258; 0,29124]
	2-***nn*** **com distância euclideana**		
ρ	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
$-0,3$	[−0,48175; −0,46820]	[−0,46879; −0,45494]	[−0,48628; −0,47290]
$-0,2$	[−0,33701; −0,31819]	[−0,32644; −0,30716]	[−0,34077; −0,32179]
$-0,1$	[−0,19484; −0,16905]	[−0,18870; −0,16386]	[−0,19629; −0,17140]
0,1	[0,12498; 0,16618]	[0,11576; 0,15542]	[0,12820; 0,17016]
0,2	[0,30586; 0,36060]	[0,29159; 0,34444]	[0,31063; 0,36602]
0,3	[0,48693; 0,57518]	[0,47330; 0,55801]	[0,49142; 0,58080]
$-0,3$	[−0,37667; −0,32903]	[−0,27742; −0,21097]	[−0,44667; −0,40059]
$-0,2$	[−0,25170; −0,19338]	[−0,18608; −0,11551]	[−0,30267; −0,23846]
$-0,1$	[−0,16321; −0,09958]	[−0,13972; −0,06334]	[−0,18949; −0,11860]
0,1	[0,04002; 0,11131]	[−0,02596; 0,05423]	[0,08466; 0,16861]
0,2	[0,17613; 0,24994]	[0,06033; 0,14028]	[0,25296; 0,33996]
0,3	[0,35431; 0,43787]	[0,20267; 0,28259]	[0,44037; 0,53968]
$-0,3$	[−0,24606; −0,17167]	[−0,17204; −0,08219]	[−0,35212; −0,26947]
$-0,2$	[−0,17603; −0,09785]	[−0,13979; −0,04674]	[−0,24681; −0,14989]
$-0,1$	[−0,12809; −0,04604]	[−0,11992; −0,02504]	[−0,16163; −0,05697]
0,1	[−0,03854; 0,04996]	[−0,08195; 0,01368]	[0,01275; 0,12603]
0,2	[0,03187; 0,12177]	[−0,05095; 0,04409]	[0,13358; 0,25184]
0,3	[0,15795; 0,24809]	[0,01296; 0,10871]	[0,30991; 0,42407]

Fonte: Elaborada pela autora.

Tabela 14 – Intervalos de confiança para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, com diferentes variabilidades, utilizando o k-*nn* com a distância de Hausdorff para o cálculo da matriz $\mathbf{\Lambda}$.

Intervalos de confiança (5% de erro)			
1-*nn* com distância de Hausdorff			
ρ	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
$-0,3$	[−0,53784; −0,51589]	[−0,52089; −0,49839]	[−0,54434; −0,52180]
$-0,2$	[−0,39515; −0,36895]	[−0,38213; −0,35535]	[−0,40002; −0,37304]
$-0,1$	[−0,24657; −0,21650]	[−0,23993; −0,20948]	[−0,24902; −0,21859]
$0,1$	[0,05650; 0,09389]	[0,04793; 0,08551]	[0,05955; 0,09737]
$0,2$	[0,20782; 0,24962]	[0,19303; 0,23400]	[0,21270; 0,25538]
$0,3$	[0,35272; 0,40185]	[0,33421; 0,38170]	[0,35927; 0,40941]
$-0,3$	[−0,41710; −0,34559]	[−0,31020; −0,21606]	[−0,50868; −0,43037]
$-0,2$	[−0,29884; −0,21793]	[−0,22385; −0,12654]	[−0,36289; −0,26847]
$-0,1$	[−0,21977; −0,13668]	[−0,19128; −0,09074]	[−0,26080; −0,16161]
$0,1$	[−0,03465; 0,05367]	[−0,08633; 0,01657]	[−0,00627; 0,10110]
$0,2$	[0,06316; 0,15386]	[−0,02928; 0,07305]	[0,12571; 0,23216]
$0,3$	[0,16995; 0,26083]	[0,03795; 0,13699]	[0,26308; 0,36547]
$-0,3$	[−0,27401; −0,17451]	[−0,20934;−0,08782]	[−0,40178; −0,26712]
$-0,2$	[−0,22090; −0,11834]	[−0,18628; −0,06290]	[−0,31289; −0,16640]
$-0,1$	[−0,17562; −0,07047]	[−0,16814; −0,04257]	[−0,22983; −0,07906]
0,1	[−0,09499; 0,01482]	[−0,13247; −0,00798]	[−0,07222; 0,08311]
0,2	[−0,04976; 0,05838]	[−0,11310; 0,01087]	[0,01372; 0,17104]
0,3	[0,00402; 0,11395]	[−0,08919; 0,03414]	[0,10793; 0,26420]
2-*nn* com distância de Hausdorff			
ρ	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
$-0,3$	[−0,49832; −0,48776]	[−0,48550; −0,47438]	[−0,50300; −0,49247]
$-0,2$	[−0,35829; −0,34327]	[−0,34685; −0,33182]	[−0,36228; −0,34727]
$-0,1$	[−0,21873; −0,19902]	[−0,21204; −0,19295]	[−0,22108; −0,20116]
$0,1$	[0,09723; 0,12435]	[0,08867; 0,11527]	[0,10019; 0,12757]
$0,2$	[0,28023; 0,31737]	[0,26669; 0,30144]	[0,28496; 0,32310]
$0,3$	[0,47220; 0,54062]	[0,45805; 0,52332]	[0,477285; 0,54655]
$-0,3$	[−0,38734; −0,34221]	[−0,28423; −0,22292]	[−0,46269; −0,41309]
$-0,2$	[−0,27242; −0,21823]	[−0,20385; −0,13535]	[−0,33147; −0,26956]
$-0,1$	[−0,17766; −0,11590]	[−0,14686; −0,07431]	[−0,21035; −0,13811]
0,1	[0,01722; 0,08553]	[−0,03803; 0,03845]	[0,05218; 0,13781]
0,2	[0,14828; 0,21861]	[0,04283; 0,11701]	[0,22067; 0,30372]
0,3	[0,33564; 0,40566]	[0,18401; 0,25546]	[0,42146; 0,50466]
$-0,3$	[−0,24798; −0,17829]	[−0,17485; −0,08842]	[−0,36295; −0,27012]
$-0,2$	[−0,17940; −0,10466]	[−0,14323; −0,05307]	[−0,25672; −0,15196]
$-0,1$	[−0,13249; −0,05493]	[−0,12287; −0,03131]	[−0,17353; −0,06086]
0,1	[−0,04544; 0,03543]	[−0,08661; 0,00655]	[−0,00723; 0,11365]
0,2	[0,02012; 0,10125]	[−0,05796; 0,03591]	[0,10728; 0,23155]
0,3	[0,13847; 0,22346]	[0,00283; 0,09481]	[0,28523; 0,39823]

Fonte: Elaborada pela autora.

3.3 DADOS INTERVALARES REAIS

Nessa seção investigamos o comportamento dos índices propostos a partir de duas aplicações envolvendo dados reais intervalares. Para a análise, foram utilizados o índice de Moran clássico e o índice de autocorrelação espacial intervalar com diferentes representações. Dois conjuntos de dados clássicos foram agregados para formar conjuntos de dados simbólicos.

3.3.1 Aplicação I

A presente subseção ilustra a aplicação de um conjunto de dados clássicos que é agregado para uma variável simbólica do tipo de intervalo. Esse conjunto de dados apresenta informações sobre a localização geográfica. Assim, estamos interessados na análise exploratória de dados de contagens que são espacialmente 'rotulados' (ou seja, são indexados por uma localização geográfica). Aplicamos os índices descritos acima ao conjunto de dados *rent99* do pacote *gamlss.data* do R, como forma de ilustrar seu uso prático (RIGBY et al., 2019). Estes dados referem-se aos preços de aluguéis (em EURO) na cidade de Munique, que consiste em 3082 observações para 411 regiões da cidade. Para facilitar a visualização dessas regiões, foi construído um mapa, apresentado na Figura 8.

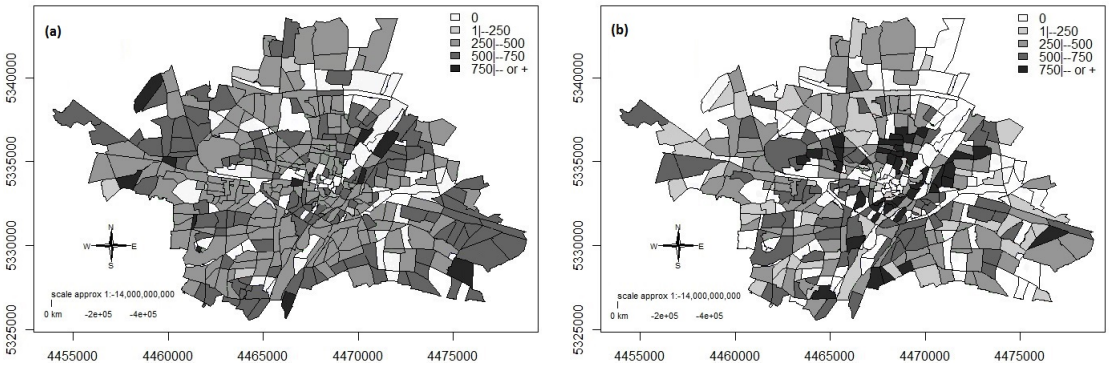
De acordo com a Tabela 15, a variável preço de aluguel para a cidade de Munique é agregada por uma variável simbólica do tipo intervalo. Para este procedimento de agregação, o código de identificação (ID) de cada região é tomado como referência. Assim, ao invés de termos diversos valores para cada região, teremos um intervalo. Por exemplo, para a região R_2 cujo ID é 2, os preços de aluguéis assumem valores na faixa [61,538; 1060,564]. Outra abordagem consiste em considerar a variabilidade das coordenadas geográficas, nas quais são considerados os valores mínimo e máximo da latitude e longitude. Assim, diferentemente dos dados clássicos, para os quais cada ponto de dados consiste em um único valor (categórico ou quantitativo), os dados simbólicos contêm variação interna. Desta forma, estudaremos dados que levam em consideração a variação interna existente e a dependência espacial.

Tabela 15 – Conjunto de dados simbólicos intervalar para o preço de aluguéis na cidade de Munique.

Região	Centro da Latitude	Centro da Longitude	Latitude intervalar	Longitude intervalar	Preços de aluguéis intervalar
R_1	4469304	5333696	[4468657; 4469111]	[5333510; 5334138]	[187,949; 701,896]
R_2	4469595	5333550	[4468531; 4468668]	[5333515; 5333927]	[61,538; 1060,564]
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R_{410}	446927	5333037	[4462452; 4463520]	[5332050; 5332705]	[220,410; 1181,589]
R_{411}	4469643	5333961	[4461788; 4462107]	[5333291; 5333787]	[461,128; 914,051]

Fonte: Elaborada pela autora.

Figura 14 – Média (a) e Amplitude (b) dos preços de aluguéis na cidade de Munique.



Fonte: Elaborada pela autora.

Podemos ver na Fig. 14(a) a distribuição dos aluguéis na região de Munique. Os pontos mais escuros do mapa correspondem as regiões com os preços médios mais altos. Para verificar a variabilidade dentro das regiões, a distribuição espacial da amplitude é observada na Fig. 14(b). A média intervalar para o preço do aluguel e o desvio padrão na cidade de Munich são 421,2519 EUROS e 273,9148 EUROS, respectivamente, calculados a partir das Equações 2.2 e 2.3. A partir desse mapa, é possível observar uma grande variabilidade no preço dos aluguéis com a distribuição espacial da amplitude. Esse comportamento indica a necessidade de um índice de autocorrelação espacial, que capture a variabilidade existente nas regiões.

Para medir a dependência espacial entre valores de uma mesma variável em diferentes locais no espaço, o índice de Moran e o índice de autocorrelação espacial intervalar foram utilizados para avaliar a autocorrelação espacial. Na Tabela 16, verificamos os valores dos índices calculados com diferentes matrizes de vizinhança. Podemos ver que os índices apresentam comportamentos semelhantes na maioria dos cenários. Independentemente do critério de vizinhança utilizado ou dos índices, os valores obtidos para os índices indicaram a ausência de autocorrelação espacial.

Tabela 16 – Valores para o I, $IASi$, $IASi^*$ e $IASi^*$(amplitude=0), com a matriz **W**.

Critério de vizinhaça	**Dados clássicos**	**Dados intervalares**		
	I	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
1-centroide mais próximo	0,0644	0,1098	0,1375	0,0835
2-centroides mais próximos	0,0649	0,1114	0,1256	0,0980
Distância Chebyshev	0,1320	0,0522	$-0,0078$	0,1090
Contiguidade da rainha	0,1122	0,1675	0,1811	0,1547

Fonte: Elaborada pela autora.

A partir desses resultados, também podemos examinar o comportamento do IASi*. O seu valor é maior do que o obtido para o índice clássico, quando utilizamos o 2-centroides mais próximos como critério de vizinhança. Assim, quando utilizamos uma variável simbólica do tipo intervalar, os índices correspondentes conseguem capturar a autocorrelação espacial de forma mais sucinta, quando comparados ao caso clássico. Ou seja, a interpretação do índice é a mesma independentemente do tipo de dado utilizado.

É intuitivo pensarmos que quando lidamos com pequenas regiões, a homogeneidade tende a ser alta. Por outro lado, para regiões maiores, esperamos grande variabilidade nos dados. Ou seja, regiões maiores tendem a ser mais heterogêneas (ponto de vista geográfico). Dessa forma, dividimos a região de Munique em dois grupos (Fig. 15 (a)) para verificar se a redução da área impactará no comportamento dos índices.

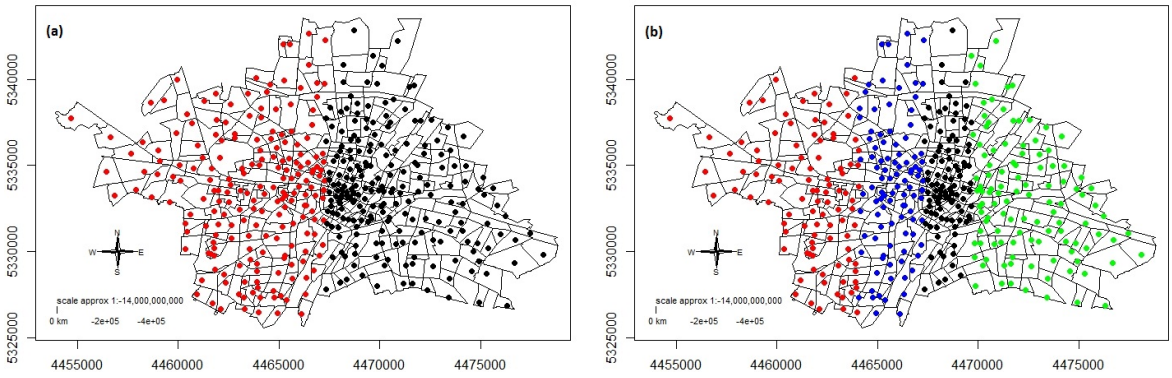
De acordo com os resultados da Tabela 17, observa-se que eles apresentaram o mesmo comportamento para autocorrelação em quase todos os cenários. Ao analisar o cenário em que é considerada a variabilidade existente dentro de cada região, observa-se que o tipo de matriz de ponderação utilizada resulta em diferentes interpretações para o tipo de autocorrelação existente. Dentre os critérios de vizinhança utilizados, o critério de contiguidade da rainha foi o que apresentou ausência de autocorrelação espacial sem alterar a interpretação do tipo de autocorrelação.

Tabela 17 – Valores obtidos para os índices I, $IASi$, $IASi^*$ e $IASi^*$(amplitude=0), quando a cidade é dividida em duas regiões (Grupos), utilizando a matriz **W**.

Critério de vizinhaça	**Grupos**	**Dados clássicos**	**Dados intervalares**		
		I	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
1-*nc*	G_1	$-0,0260$	$-0,0402$	$-0,0382$	$-0,0382$
	G_2	0,1103	0,1815	0,2432	$-0,0382$
2-*nc*	G_1	$-0,0108$	0,0147	0,0482	0,0482
	G_2	0,0975	0,1570	0,1818	0,1818
Distância de Chebyshev	G_1	$-0,0058$	$-0,0257$	$-0,1019$	$-0,1019$
	G_2	0,2805	0,1295	$-0,2089$	$-0,2089$
Contiguidade da rainha	G_1	0,0337	0,0296	0,0387	0,0387
	G_2	0,0388	0,0414	0,0636	0,0636

Fonte: Elaborada pela autora.

Figura 15 – Cidade de Munique dividida em dois grupos(a), e em quatro grupos (b).



Fonte: Elaborada pela autora.

Quando analisamos os resultados dos índices de forma conjunta, utilizando os mesmos critérios, identificamos que na maioria dos casos teremos a mesma interpretação, ou seja, se a autocorrelação for positiva no cenário em que a variável observada é clássica, o mesmo ocorre no cenário de intervalo.

Para verificar se o número de áreas em uma região impactava o valor do índice tanto no cenário onde a variável observada era clássica quanto simbólica, o mapa foi dividido em quatro partes distintas (Fig. 15 (b)). De acordo com a Tabela 18, podemos analisar o comportamento do índice de Moran e dos índices de autocorrelação espacial intervalar propostos. Em todos os cenários apresentados, observamos que a autocorrelação espacial existente é fraca.

Quando levamos em consideração a variabilidade existente em cada região, ou seja, consideramos a variável simbólica do tipo intervalo, observamos que o IASi, em geral, apresenta o mesmo comportamento do índice clássico. Para verificar essa característica, considere o índice tanto no cenário clássico quanto no intervalar. Seu valor indica uma autocorrelação fraca na maioria dos grupos.

Tabela 18 – Valores para o I, $IASi$, $IASi^*$ e $IASi^*$(amplitude$= 0$), quando a cidade é dividida em quatro grupos, utilizando a matriz de pesos **W**.

Critério de vizinhaça	**Grupos**	**Dados clássicos**	**Dados intervalares**		
		I	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
1-centroide mais próximo	G_1	0,0811	0,0290	$-0,0072$	$0,0561$
	G_2	$-0,1407$	$-0,0750$	$-0,0419$	$-0,1053$
	G_3	0,0974	0,1469	0,1895	0,0972
	G_4	0,1360	0,2179	0,2213	0,2150
2-centroides mais próximos	G_1	0,0105	$-0,0271$	$-0,0355$	$-0,0208$
	G_2	$-0,1005$	$-0,0165$	0,0115	$-0,0422$
	G_3	0,1376	0,1675	0,1767	0,1567
	G_4	0,0896	0,1578	0,1533	0,1616
Distância de Chebyshev	G_1	$-0,1191$	$-0,0841$	$-0,0251$	$-0,1285$
	G_2	0,0866	0,0452	0,0254	0,0634
	G_3	$-0,0721$	$0,0423$	0,0496	0,0338
	G_4	0,1320	0,1344	0,0860	0,1762
Contiguidade da rainha	G_1	$-0,0069$	0,0049	0,0147	$-0,0025$
	G_2	0,0004	$-0,0058$	$-0,0184$	0,0056
	G_3	0,0205	0,0320	0,0456	0,0162
	G_4	$-0,0174$	$-0,0463$	$-0,0681$	$-0,0276$

Fonte: Elaborada pela autora.

Para entender o comportamento do índice quando, tanto a variável de interesse quanto as coordenadas geográficas da região são consideradas variáveis simbólicas intervalares, os índices de autocorrelação espacial intervalares foram calculados usando a matriz de vizinhança $\mathbf{\Lambda}$. Em geral, os valores encontrados são negativos, indicando uma possível dissimilaridade entre as áreas.

Tabela 19 – Valores para o $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$, utilizando uma distância intervalar para o cálculo da matriz $\mathbf{\Lambda}$.

Critério de vizinhança	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
1-*nn* com distância de Gowda-Diday	$-0,0349$	$-0,0724$	0,0006
1-*nn* com distância de City block	$-0,0899$	$-0,0637$	$-0,1147$
1-*nn* com distância Euclideana	$-0,0700$	$-0,0602$	$-0,0793$
1-*nn* com distância de Hausdorff	$-0,1194$	$-0,0994$	$-0,1384$
2-*nn* com distância Gowda-Diday	$-0,0001$	$-0,0167$	0,0155
2-*nn* com distância de City block	$-0,0818$	$-0,0627$	$-0,0999$
2-*nn* com distância Euclideana	$-0,0864$	$-0,0766$	$-0,0957$
2-*nn* com distância Hausdorff	$-0,0800$	$-0,0528$	$-0,1057$

Fonte: Elaborada pela autora.

3.3.2 Aplicação II

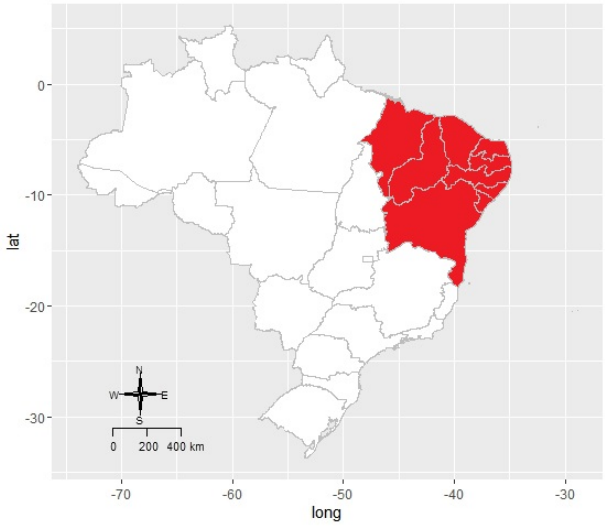
A Doença do coronavírus-19 (COVID-19) é causada pelo vírus SARS-CoV-2, que apresenta um quadro clínico que varia de infecções assintomáticas a graves condições respiratórias. De acordo com a Organização Mundial da Saúde (WORLD HEALTH ORGANIZATION, 2020), a

maioria dos pacientes com COVID-19 (cerca de 80%) pode ser assintomática e cerca de 20% dos casos podem necessitar de cuidados hospitalares.

Para mapear os dados sobre a doença, são utilizadas metodologias espaciais. Aqui, a ideia é entender o comportamento espacial do número de casos por COVID-19 no contexto de dados simbólicos intervalares. Análises estatísticas descritivas e espaciais são realizadas para dados intervalares. Para isto, usamos a base cartográfica do Nordeste brasileiro, obtida no site do Instituto Brasileiro de Geografia e Estatística. Tal base contém informações sobre 9 estados (ver Fig. 16) e 1794 municípios. O processo de agregação de vários municípios resultou em 188 microrregiões do Nordeste. Para cada região resultante do processo de agregação, os valores mínimos e máximos foram utilizados para criar os respectivos intervalos. Adota-se a variável simbólica do tipo intervalo para descrever o fenômeno, referente ao período de 25 de fevereiro à 26 de maio de 2020, no qual foram notificados 128389 casos da doença no Nordeste do Brasil.

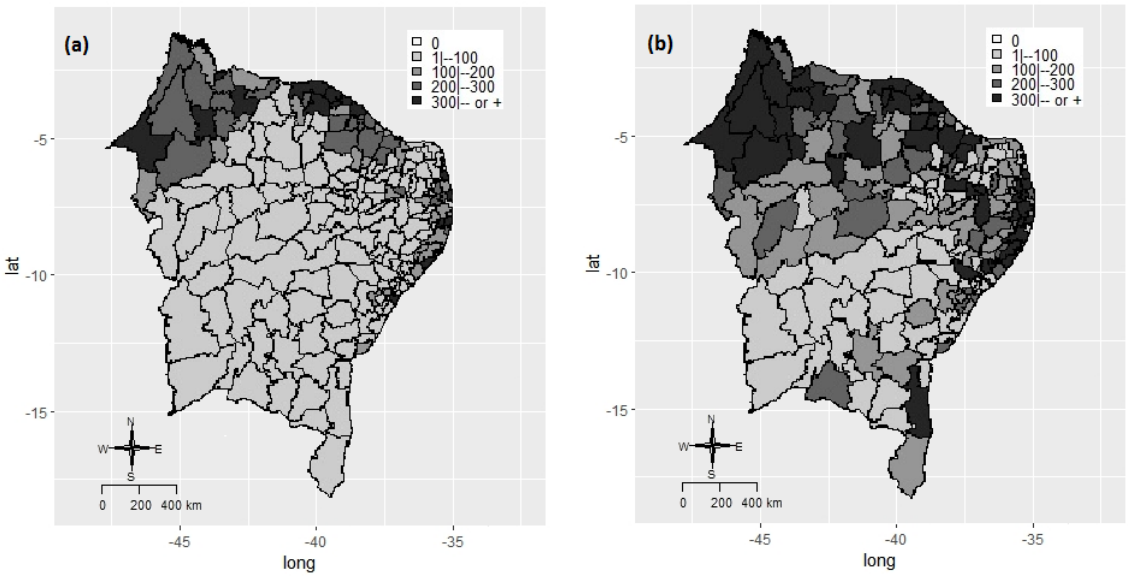
A maior incidência foi observada no município de Pedreiras, localizado no estado do Maranhão, com 1142 casos por 100 mil habitantes. Em relação às microrregiões, a maior incidência de casos da doença foi localizada na microrregião do Médio Mearim, no estado do Maranhão. A média e o desvio padrão da variável simbólica intervalar para este estado são calculados seguindo Billard e Diday (2006c), sendo dadas por 153,5729 e 186,885, respectivamente.

Figura 16 – Mapa do Brasil com destaque para a região Nordeste.



Fonte: Elaborada pela autora.

Figura 17 – Distribuição espacial da taxa de incidência de dados clássicos (a) e taxa de incidência obtida a partir de dados intervalares (b), para COVID-19 nas microrregiões do Nordeste do Brasil.

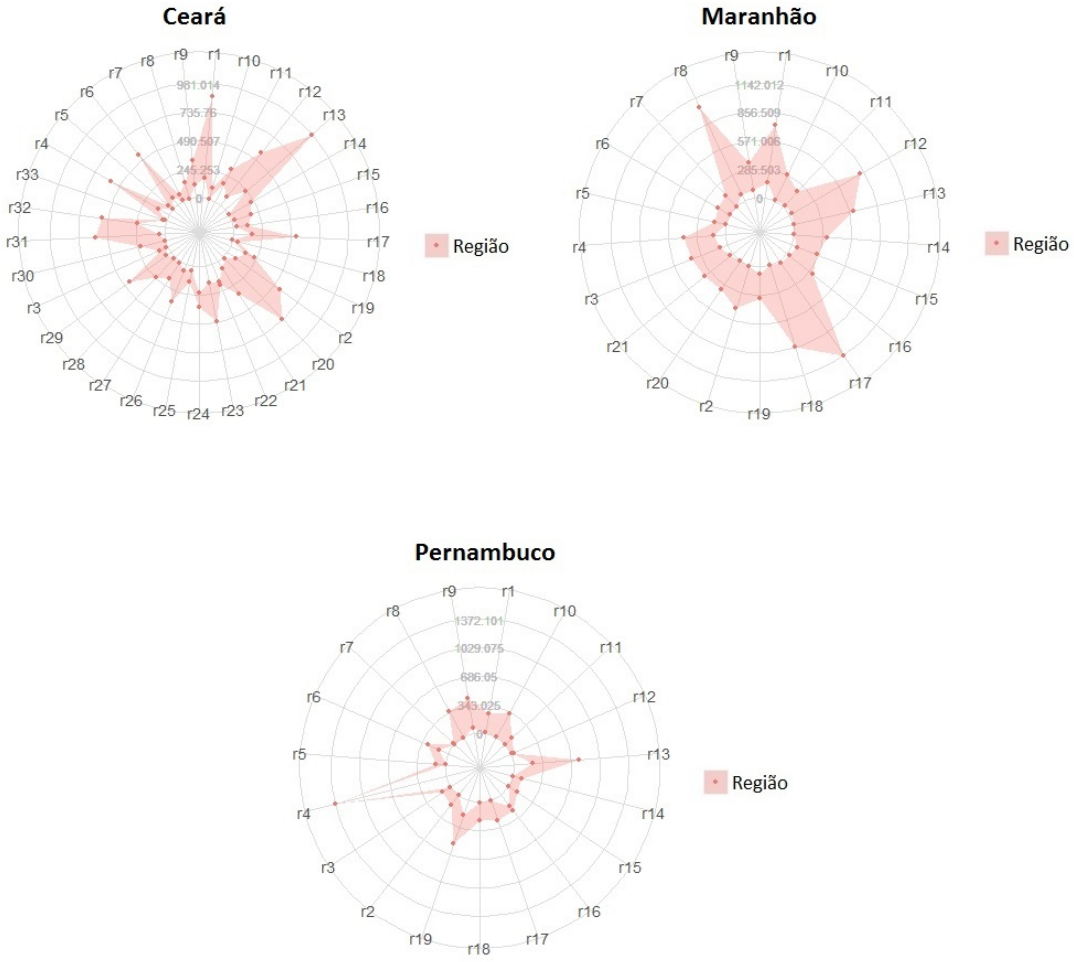


Fonte: Elaborada pela autora.

A distribuição espacial da incidência média de casos de COVID-19 por microrregião pode ser observada na Figura 17(a). Como esperado, uma alta concentração pode ser observada em regiões com alta densidade populacional. Para observar a variabilidade dentro de cada microrregião, um mapa da amplitude foi traçado na Fig. 17(b), em que os valores máximo e mínimo para cada região foram usados. Mais especificamente, tomamos a diferença entre os valores máximos e mínimos correspondentes a cada região. Ao analisar a distribuição espacial desta variável, é possível observar uma grande variabilidade nas microrregiões. Ao comparar a Fig. 17(a) com a Fig. 17(b), é possível observar claramente que a variabilidade dentro de cada microrregião não é levada em consideração em 17(a).

Para entender a variabilidade presente nas microrregiões, a Fig. 18 exibe a dispersão dos dados para alguns estados da região Nordeste. De acordo com este mapa, embora cada estado tenha comportamento distinto, é possível identificar que existem microrregiões mais críticas do que outras, que apresentam a maior variabilidade na incidência de casos. Além disso, os estados têm comportamentos diferentes no que diz respeito ao comportamento dos intervalos.

Figura 18 – Dados intervalares da COVID-19 para os estados: Ceará, Maranhão e Pernambuco.



Fonte: Elaborada pela autora.

Buscando indentificar a existência ou não de autocorrelação espacial, os índices de autocorrelação espacial foram aplicados aos dados de COVID-19, conforme mostrado na Tabela 20. Os valores obtidos nessa tabela, são calculados usando a matriz de ponderação **W**. O I indica uma autocorrelação espacial positiva, confirmando uma dependência espacial da distribuição da doença. No contexto dos dados simbólicos, os índices indicam uma autocorrelação espacial fraca em todas as representações. Quando o critério de vizinhança é calculado usando a distância de Chebyshev, para o índice clássico, não há dependência espacial da variável observada. Uma possível justificativa para isto se dá pelo fato de que o critério de Chebyshev considera apenas distâncias em linha reta entre os pontos centrais de cada região.

Quando consideramos as coordenadas geográficas como dados de intervalo, ou seja, quando os valores mínimo e máximo da latitude e longitude de cada região são utilizados, os índices

Tabela 20 – Valores do $IASi$, $IASi^*$ e $IASi^*(amplitude = 0)$ para dados da COVID-19, utilizando a matriz **W**.

Critério de vizinhaça	**Dados clássicos**	**Dados intervalares**		
	I	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
1-centroide mais próximo	0,5313	0,2683	0,2460	0,2884
2-centroides mais próximos	0,5255	0,2675	0,2594	0,2749
Distância de Chebyshev	0,1348	0,0616	0,0590	0,0640
Contiguidade da rainha	0,5123	0,2689	0,2642	0,2731

Fonte: Elaborada pela autora.

IASi, IASi* e IASi*(amplitude $= 0$) apresentam comportamentos diferentes dependendo do critério de distância utilizado. Por exemplo, quando adotamos o critério de 1-centroide mais próximo com a distância de Gowda-Diday, não há autocorrelação. Diferentemente, o critério com 1-centroide mais próximo para as demais distâncias de intervalo indicam uma autocorrelação espacial fraca. Nesse sentido, podemos concluir que há uma semelhança fraca entre as microrregiões, quando observamos a incidência de COVID-19. Quando o número de vizinhos aumenta no critério K-vizinhos mais próximos, o valor do índice diminui para todas as representações de dados de intervalo.

Tabela 21 – Índices de autocorrelação espacial para dados de COVID-19, utilizando a matriz de pesos $\mathbf{\Lambda}$.

Critério de vizinhança	**IASi**	**IASi***	**IASi**$^*(amplitude = 0)$
1-*nn* com distância Gowda-Diday	0,0006	0,0366	$-$0,0560
1-*nn* com distância City block	0,2116	0,1892	0,2469
1-*nn* com distância Euclideana	0,1629	0,1266	0,2202
1-*nn* com distância Hausdorff	0,1634	0,1167	0,2368
2-*nn* com distância Gowda-Diday	$-$0,0161	0,0098	$-$0,0569
2-*nn* com distância City block	0,1360	0,1158	0,1678
2-*nn* com distância Euclideana	0,1083	0,0914	0,1349
2-*nn* com distância Hausdorff	0,1203	0,0727	0,1951

Fonte: Elaborada pela autora.

3.4 CONSIDERAÇÕES FINAIS

Ao longo desse capítulo, introduzimos um novo índice de autocorrelação para dados intervalares na literatura de SDA. Este índice é uma extensão do índice de Moran para dados de intervalo, em que diferentes formas para o cálculo da matriz de vizinhança são abordadas. A ideia inicial para o cálculo desse índice consistiu em considerar a variável do tipo intervalar, bem como os valores centrais para a representação da latitute e longitude de cada região analisada. Também discutimos uma outra abordagem, em que empregamos a variável sim-

bólica do tipo intervalar e a informação da latitude e longitude descrita por um intervalo. Nesse caso, distâncias intervalares foram usadas para a identificação da região de vizinhança. Com esta abordagem, foi possível acomodar a variabilidade presente na variável intervalar e a variabilidade presente nas coordenadas geográficas de cada região de interesse.

Posteriomente, realizamos estudos através de simulações computacionais para ilustrar o comportamento da abordagem proposta em diferentes cenários. Os resultados demonstraram que o índice de autocorrelação espacial introduzido foi capaz de capturar a autocorrelação existente nos dados. Isto se justifica pelo fato de que, na presença da dados com e sem autocorrelação espacial, os índices intervalares apresentaram resultados concordantes com a literatura para os índices clássicos. Além disso, os resultados obtidos para conjuntos de dados reais também ilustraram a importância do índice proposto em relação aos diferentes critérios para o cálculo da matriz de pesos a partir de distâncias para dados intervalares. De posse desses resultados, foi possível notar a aplicabilidade da presente proposta sob diferentes aspectos.

4 REGRESSÃO LINEAR PARA DADOS ESPACIAIS NO CONTEXTO INTERVALAR

Modelos de regressão são ferramentas comumente empregadas para investigar a relação entre duas ou mais variáveis. O valor de uma das variáveis pode ser estimado em função das demais. No contexto da análise de dados espaciais, frequentemente os dados apresentam autocorrelação, resultando em valores elevados na significância dos parâmetros. Em cenários envolvendo autocorrelação espacial, é essencial propor modelos que possam incorporar a estrutura espacial dos dados (Ver, (BANERJEE; CARLIN; GELFAND, 2014)).

A dependência entre observações em diferentes locais é uma característica inerente aos dados espaciais, e existem várias maneiras de introduzir os efeitos dos fenômenos espaciais nos modelos de regressão. O mais simples é denominado 'modelo com efeitos espaciais globais', que busca capturar a estrutura de correlação espacial por meio de um parâmetro incluído no modelo de regressão. Um exemplo de modelo que leva em conta os efeitos espaciais é o Autoregressivo Condicional (BESAG, 1974). Normalmente, esse modelo é empregado para variáveis clássicas e considera os efeitos espaciais como um tipo de ruído, ou equivalentemente, como um fator a ser removido.

Dentre vários modelos de regressão para dados simbólicos descritos na literatura, a ausência de modelos para lidar com dados no contexto de dados georreferenciados é evidente na literatura de SDA, demandando atenção para o desenvolvimento de metodologias inovadoras. A seguir, apresentamos um modelo de regressão intervalar que considera a variabilidade na região e a dependência espacial existente no contexto de dados simbólicos intervalares. O modelo apresentado neste capítulo é inédito na literatura.

Este capítulo está organizado da seguinte forma: Na Seção 4.1, introduzimos uma notação para lidar com dados espaciais simbólicos. Na Seção 4.2, propomos um modelo de regressão espacial simbólico intervalar usando as representações centro e amplitude, e mínimo e máximo. A seção 4.3 descreve a avaliação experimental com dados intervalares sintéticos. A Seção 4.4 mostra duas aplicações com dados reais. A primeira está relacionada aos dados de mortalidade infantil no estado de Pernambuco, enquanto a segunda refere-se a informações sobre a renda per capita no nordeste do Brasil. Finalmente, a Seção 4.5 apresenta as considerações finais sobre as abordagens utilizadas.

4.1 DADOS ESPACIAIS DE INTERVALO SIMBÓLICO

Nessa seção, o foco está na análise de dados onde a variável intervalar se relaciona com geo-objetos. Vamos considerar um conjunto de dados $E = \{1, \cdots, n\}$ de n elementos descritos por p variáveis de intervalo $Z_1 = [a_1; b_1], Z_2 = [a_2; b_2], \ldots, Z_p = [a_p; b_p]$ e a localização da área $S(i)$. De acordo com Haining e Haining (2003), podemos representar esses dados por uma matriz de dados espaciais.

No contexto da análise de dados simbólicos, essa representação é dada por

$$
\begin{array}{cccccc}
 & p \text{ variáveis intervalares} & & & \text{Localização} & \\
 & Z_1 & Z_2 & \ldots & Z_p & S(i) \\
\left[\begin{array}{cccccc}
 & Z_{11} = [a_{11}, b_{11}] & Z_{12} = [a_{12}, b_{12}] & \ldots & Z_{1p} = [a_{1p}, b_{1p}] & \mathsf{S(1)} \\
 & Z_{21} = [a_{21}, b_{21}] & Z_{22} = [a_{22}, b_{22}] & \ldots & Z_{2k} = [a_{2p}, b_{2p}] & \mathsf{S(2)} \\
 & \vdots & \vdots & & \vdots & \vdots \\
 & Z_{n1} = [a_{n1}, b_{n1}] & Z_{n2} = [a_{n2}, b_{n2}] & \ldots & Z_{np} = [a_{np}, b_{np}] & \mathsf{S(n)}
\end{array}\right] & \begin{array}{c} \text{Região 1} \\ \text{Região 2} \\ \vdots \\ \text{Região } n \end{array}
\end{array}
$$

que pode ser representado por

$$\{Z_{i1}, Z_{i2}, \cdots, Z_{ip} \mid \mathsf{S(i)}\}_{i=1,\cdots,n} \tag{4.1}$$

onde S(i) $(i = 1, \ldots, n)$ representa um vetor de localização espacial da i-ésima região. Como estamos interessados apenas no espaço bidimensional, $\mathbf{s}_i$ correspondente ao par de coordenadas geográficas, s_{i1} e s_{i2}, que são valores centrais de latitude e longitude, respectivamente.

Considere parte de um conjunto de dados clássicos reais sobre a incidência de Corona Virus (COVID-19) exibido na Tabela 22. Esse conjunto de dados é descrito pelas seguintes variáveis: cidade, número de casos, latitude, longitude, número de óbitos e população

A partir das variáveis número de óbitos e número de casos, podemos construir as variáveis espaciais intervalares Z_1 e Z_2. Fixada uma região i, os valores mínimos e máximos para essa região podem ser representados por uma variável intervalar [112, 400] e [950, 1500], respectivamente. Com base na representação expressa em (4.1), esta região é representada por $\{[112, 400], [950, 1500], \mid \mathsf{S(i)} = (-6, 49; 38, 45)\}$.

As informações georreferenciadas definem não apenas quais pares de áreas são adjacentes, mas também podem quantificar a proximidade dessa adjacência. Essas informações são

Tabela 22 – Conjunto de dados de COVID-19.

Cidade	**Região**	**Latitude**	**Longitude**	**Número de casos**	**Número de óbitos**	**População**
Cidade_1	R_1	-6.46	-38.12	950	112	69997
Cidade_2	R_1	-6.80	-38.00	1122	225	407472
Cidade_3	R_1	-6.83	-38.40	1500	400	5088312
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
Cidade_{223}	R_{23}	-7.34	-34.47	1720	155	4059905

Fonte: Elaborada pela autora.

necessárias para especificar muitos modelos estatísticos espaciais, como modelos de regressão espacial. De acordo com Cressie e Chan (1989), os pesos podem ser baseados na distância. Para gerar o grau de conectividade entre as regiões, utilizamos distâncias calculadas entre os centroides de cada região. Cada região é identificada por suas coordenadas de latitude e longitude, descritas por valores centrais. Assim, para gerar a matriz de pesos utilizamos os seguintes critérios: K-vizinhos mais próximos, com $K = 1, 2$ (NEGREIROS et al., 2010); Distância de Chebyshev e contiguidade da rainha (FISCHER; WANG, 2011).

4.2 MODELO DE REGRESSÃO ESPACIAL PARA DADOS DE VALOR SIMBÓLICO INTERVALAR

Esta seção apresenta um modelo de regressão espacial para dados intervalares em que as abordagens espacial e simbólica são analisadas conjuntamente. A seguir, propomos duas abordagens diferentes para ajustar modelos de regressão espacial na estrutura de dados intervalares. Seja Ω, um conjunto de dados de n regiões descritas pela variável resposta com valor de intervalo Y e p variáveis preditoras com valor de intervalo $(X_1, \ldots, X_p)$. Cada elemento i de Ω é representado como um vetor de características $\mathcal{F}_i = (\mathbf{x}_i, y_i, s(i))$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ onde $x_{ij} = [a_{ij}, b_{ij}] \in \mathfrak{I} = \{[a, b] : a, b \in \Re, a \leq b\} (j = 1, \ldots, p)$, $y_i = [\alpha_i, \lambda_i] \in \mathfrak{T}$ (BOCK; DIDAY, 1999) e S(i) $= (s_{i1}, s_{i2})$ representa o vetor latitude e longitude da região i.

A regressão espacial para dados de intervalo é dada por:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.2}$$

com

$$\boldsymbol{\epsilon} = \Lambda \mathbf{W}_{\epsilon} + \xi.$$

em que $\mathbf{Y}$ é o vetor $n \times 1$ da variável resposta; $\mathbf{X}$ é a matriz $n \times p$ de $X_1, \cdots, X_p$ variáveis

explicativas; $\boldsymbol{\beta}$ é vetor de $p \times 1$ de coeficientes da regressão a ser estimado; A componente do erro com efeitos espaciais é expressa por W_ϵ; Λ representa o coeficiente autorregressivo e ξ é o vetor $n \times 1$ da componente do erro com média zero, variância constante e não correlacionada.

Uma característica crucial desse modelo, é que ele parte do princípio de que o processo espacial subjacente aos dados analisados é estacionário (MONTEIRO et al., 2004). Assim, padrões de autocorrelação espacial existentes nos dados podem ser capturados em um único parâmetro.

4.2.1 Representação Mínimo e Máximo (Min-Max)

Em relação aos valores mínimo e máximo dos intervalos, temos:

- o vetor $\mathbf{Y}$ e a matriz $\mathbf{X}$ são definidos por: $\mathbf{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4})$, com $\boldsymbol{X_1} = (\mathbf{1}_n^\top, \mathbf{0}_n^\top)^\top$, $\boldsymbol{X_2} = (\boldsymbol{x}_L^{,}\mathbf{0}_n^\top)^\top$ $\boldsymbol{X_3} = (\mathbf{0}_n^\top, \mathbf{1}_n^\top)^\top$, e $\boldsymbol{X_4} = (\mathbf{0}_n^\top, \boldsymbol{x}_U^\top)^\top$ onde $\boldsymbol{x}_L = (a_{1j}, \ldots, a_{nj})^\top$ $\boldsymbol{x}_U = (b_{1j}, \ldots, b_{nj})^T$ $(j = 1, \ldots, p)$;
- $\mathbf{0}_n$ e $\mathbf{1}_n$ são vetores de zeros e uns, respectivamente;
- $\boldsymbol{\beta} = (\beta_0^L, \beta_1^L, \ldots, \beta_p^L, \beta_0^U, \beta_1^U, \ldots, \beta_p^U)^\top$ é o vetor de parâmetros;
- $\boldsymbol{\xi} = (\boldsymbol{\xi}^L, \boldsymbol{\xi}^U)^\top$ é o vetor de erros com $\boldsymbol{\xi}^L = (\xi_1^L, \ldots, \xi_n^L)^\top$ e $\boldsymbol{\xi}^U = (\xi_1^U, \ldots, \xi_n^U)^\top$;
- $\mathbf{W}_\epsilon = (\mathbf{W}^L, \mathbf{W}^U)$ é o vetor de erros com efeitos espaciais $\mathbf{W}^L = (W_1^L, W_2^L, \ldots, W_n^L)$ e $\mathbf{W}^U = (W_1^U, \ldots, W_n^U)$;
- Λ é um coeficiente autorregressivo.

Os coeficientes são estimados pelo método de máxima verossimilhança. Assim, para uma nova região descrita por $\mathbf{z} = (\mathbf{x}, y, s)$, onde $\mathbf{x} = (x_1, \ldots, x_p)$ com $x_j = [a_j, b_j]$ $(j = 1, \ldots, p)$ e $s = (s(1), s(2))$. O intervalo predito $\hat{y} = \left[\hat{\alpha}, \hat{\lambda}\right]$ é o seguinte:

$$\hat{\alpha} = \left(\mathbf{x}^L\right)^\top \hat{\boldsymbol{\beta}}_L \quad \text{e} \quad \hat{\lambda} = \left(\mathbf{x}^U\right)^\top \hat{\boldsymbol{\beta}}_{\mathrm{U}}, \tag{4.3}$$

sendo $\mathbf{x}^L = (a_1, \ldots, a_p)$, $\mathbf{x}^U = (b_1, \ldots, b_p)$, $\boldsymbol{\beta}_L = (\beta_0^L, \beta_1^L, \ldots, \beta_p^L)$ e $\boldsymbol{\beta}_U = (\beta_0^U, \beta_1^U, \ldots, \beta_p^U)$.

4.2.2 Representação Centro e Amplitude (Cen; Amp)

Em relação ao centro e amplitude dos intervalos, temos

- o vetor $\mathbf{Y}$ e a matriz $\mathbf{X}$ são definidos como: $\mathbf{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4})$ com $\boldsymbol{X_1} = (\mathbf{1}_n^\top, \mathbf{0}_n^\top)^\top$, $\boldsymbol{X_2} = (\boldsymbol{x}_C^\top, \mathbf{0}_n^\top)^\top$ $\boldsymbol{X_3} = (\mathbf{0}_n^\top, \mathbf{1}_n^\top)^\top$, e $\boldsymbol{X_4} = (\mathbf{0}_n^\top, \boldsymbol{x}_R^\top)^\top$ onde $\boldsymbol{x}_C = ((a_{1j}+b_{1j})/2, \ldots, (a_{nj}+b_{nj})/2)^\top$, $\boldsymbol{x}_R = ((b_{1j}-a_{1j}), \ldots, (b_{nj}-a_{nj}))^\top$ $(j = 1, \ldots, p)$;
- $\mathbf{0}_n$ e $\mathbf{1}_n$ são vetores de zeros e uns, respectivamente;
- $\boldsymbol{\beta} = (\beta_0^C, \beta_1^C, \ldots, \beta_p^C, \beta_0^R, \beta_1^R, \ldots, \beta_p^R)^\top$ é um vetor de parâmetro;
- $\boldsymbol{\xi} = (\boldsymbol{\xi}^C, \boldsymbol{\xi}^R)^\top$ é um vetor de erro com $\boldsymbol{\xi}^C = (\xi_1^C, \ldots, \xi_n^C)^\top$ e $\boldsymbol{\xi}^R = (\xi_1^R, \ldots, \xi_n^R)^\top$;
- $\mathbf{W}_\epsilon = (\mathbf{W}^C, \mathbf{W}^R)$ é um vetor de erro com efeito espacial com $\mathbf{W}^C = (W_1^C, \ldots, W_n^C)$ e $\mathbf{W}^R = (W_1^R, \ldots, W_n^R)$;
- Λ é um coeficiente autorregressivo.

Os coeficientes são estimados pelo método de máxima verossimilhança. Então, para uma nova região descrita por $\mathbf{z} = (\mathbf{x}, y, s)$, onde $\mathbf{x} = (x_1, \ldots, x_p)$ com $x_j = [a_j, b_j]$ $(j = 1, \ldots, p)$ e $s = (s(1), s(2))$. O intervalo predito $\hat{y} = \left[\hat{\alpha}, \hat{\lambda}\right]$ é o seguinte:

$$\hat{\alpha} = \hat{y}^C - \hat{y}^R/2 \quad \text{e} \quad \hat{\lambda} = \hat{y}^C + \hat{y}^R/2, \tag{4.4}$$

onde

$$\hat{y}^C = (\mathbf{x}^c)^{\mathrm{T}}\, \hat{\boldsymbol{\beta}}_C \quad \text{e} \quad \hat{y}^R = \left(\mathbf{x}^R\right)^{\mathrm{T}} \hat{\boldsymbol{\beta}}_R, \tag{4.5}$$

com

$$\left(\mathbf{x}^C\right)^{\mathrm{T}} = (1, (a_1 + b_1)/2, \ldots, (a_p + b_p)/2))\,, \tag{4.6}$$

$$\left(\mathbf{x}^R\right)^{\mathrm{T}} = (1, (b_1 - a_1), \ldots, (b_p - a_p))\,, \tag{4.7}$$

$$\hat{\boldsymbol{\beta}}_c = \left(\hat{\beta}_0^c, \hat{\beta}_1^c, \ldots, \hat{\beta}_p^c\right)^{\mathrm{T}} \text{ e } \hat{\boldsymbol{\beta}}_R = \left(\hat{\beta}_0^r, \hat{\beta}_1^r, \ldots, \hat{\beta}_p^r\right)^{\mathrm{T}}. \tag{4.8}$$

4.3 AVALIAÇÃO EXPERIMENTAL COM DADOS SINTÉTICOS INTERVALARES

O desempenho do método proposto é avaliado com experimentos em que conjuntos de dados com valores intervalares sintéticos são utilizados na simulação de Monte Carlo. Para isso,

simulamos dados espaciais e avaliamos o desempenho do modelo de regressão linear intervalar e um modelo de regressão espacial. Nesse contexto, a ideia é verificar o comportamento do modelo quando o dado analisado é do tipo intervalar e traz consigo a informação da sua localização. Assim, a proposta de Neto e Carvalho (2008) é utilizada para a representação de centro e amplitude, enquanto a proposta de Billard e Diday (2002) é empregada para a representação de mínimo e máximo. Da literatura de modelos espaciais, o modelo utiizado é o CAR.

A performance é medida pela média da magnitude relativa do erro da estimativa (MMREE). De acordo com a literatura de SDA, podemos encontrar várias abordagens para avaliar o desempenho do modelo de regressão. Exemplos dessas abordagens podem ser encontrados em Fagundes, Souza e Cysneiros (2013), Neto e Carvalho (2008). Neste capítulo, usamos a MMREE (FAGUNDES; SOUZA; CYSNEIROS, 2013) expressa por:

$$MMREE = \frac{1}{2n}\sum_{i=1}^{n}\left(\left|\frac{\alpha_i - \hat{\alpha}_i}{\alpha_i}\right| + \left|\frac{\lambda_i - \hat{\lambda}_i}{\lambda_i}\right|\right). \tag{4.9}$$

Para diferentes configurações do conjunto de dados, calculamos a MMREE. Em cada iteração da simulação de Monte Carlo, o modelo de regressão para um conjunto de dados intervalar é ajustado. Para cada 100 réplicas, a média é calculada. Três diferentes cenários de conjuntos de dados são considerados. Para cada cenário, empregamos diferentes matrizes de peso para introduzir a dependência espacial.

O algoritmo 3 inicia-se com a geração de conjuntos de dados pontuais que são sementes para cada região. Assim, para cada área, é gerado um subconjunto de valores contínuos com diferentes valores espaciais de autocorrelação. Nesse contexto, o uso de uma matriz de peso é importante para obtenção de dados com autocorrelação espacial (BIVAND et al., 2008). Os dados intervalares são obtidos para cada região (classe) a partir dos valores mínimo e máximo da distribuição normal associada a cada região.

Os critérios apresentados na seção 2.2.1 são considerados para gerar o grau de conectividade entre as regiões. Cada área é descrita por suas coordenadas de latitude e longitude, que são valores centrais.

As Tabelas 23 e 24 apresentam a média e o desvio padrão da MMREE utilizando o mínimo e máximo (Min; Max) e o centro e amplitude (Cen; Amp) dos intervalos para os modelos de regressão linear intervalar e CAR intervalar, respectivamente. De acordo com a Tabela 23, podemos verificar que o desempenho dos modelos de regressão intervalar para dados

Algoritmo 3: Simulação de Monte Carlo.

begin
  Given um mapa com regiões irregulares;
  Requer MC = 100 **for** j *tal que* $1 \leq g \leq MC$ **do**
    Generate um conjunto de dados autocorrelacionados da região $\{q_1, \ldots, q_n\}$ de acordo com uma distribuição normal com média igual a 100 e variância 36 e a matriz calculada com base no critério adotado. Cada valor contínuo está associado a uma das regiões n.
    for $i = 1 : n$ **do**
      If o cenário 1 é considerado para **gerar** um subconjunto de valores contínuos para a região R_i de acordo com uma distribuição normal com média igual a q_i e variância igual a um valor gerado a partir de uma distribuição uniforme no intervalo $[1, 10]$ (**baixo grau de concentração de dados do ponto**);
      If o cenário 2 é usada para **gerar** um subconjunto de valores contínuos para a região i de acordo com uma distribuição normal com média igual a q_i e variância igual a um valor gerado a partir de uma distribuição uniforme no intervalo $[1, 50]$ (**grau moderado de concentração de dados pontuais**);
      If o cenário 3 é usada para **gerar** um subconjunto de valores contínuos para a região i de acordo com uma distribuição normal com média igual a q_i e variância igual a um valor gerado a partir de uma distribuição uniforme no intervalo $[1, 100]$ (**alto grau de concentração de dados pontuais**). **Compute** os valores mínimo e máximo de $x_i = [a_i, b_i]$ para a região R_i;
    end
    for $i = 1 : n$ **do**
      Compute $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ com $\beta_0 = 1,5$; $\beta_1 = 2$ e ϵ_i segue uma distribuição normal com média 0 e variância 1.
    end
  end
  Define o conjunto de treinamento aleatoriamente (70% do conjunto de dados da simulação)
  Build o modelo de regressão para a representação mínimo e máximo ou centro e amplitude dos intervalos para o conjunto de dados de treino
  Compute os intervalos previstos usando o conjunto de teste (30% do conjunto de dados da simulação)
  Compute a taxa de precisão do modelo
end
Compute a média e o desvio padrão da taxa de precisão.

autocorrelacionados espaciais é melhor à medida que a autocorrelação nos dados aumenta $(\rho = 0,6)$. Esse comportamento é independente do cenário analisado. Como exemplo, considere na Tabela 23 a parametrização (Min; Max), onde o critério de 1-centroide mais próximo foi utilizado para calcular a matriz de vizinhança na geração dos dados. Quando os valores de ρ são $-0,6$ e $0,6$, os valores da MMREE são $0,0155$ e $0,0034$. Se considerarmos o critério da rainha para gerar a matriz de vizinhança, esse comportamento é semelhante ao desse modelo.

Ao comparar o cenário 1 (menor variabilidade) com o cenário 3 (maior variabilidade), observamos que para valores positivos de ρ, a parametrização mínimo e máximo apresenta melhor desempenho. No entanto, valores negativos de ρ resultam em um comportamento distinto; ou seja, à medida que a variabilidade aumenta, o modelo tem um desempenho melhor.

Ao analisar o desempenho dos modelos usando 2-vizinhos mais próximos, verificamos que o desempenho é afetado pelo valor da autocorrelação espacial. Para valores positivos e pouca variabilidade, encontramos um melhor desempenho do modelo. Ao considerar a distância de Chebyshev ou o critério da rainha na geração de dados, observamos que modelos com alta autocorrelação espacial e menor variabilidade nos dados tendem a apresentar melhores resultados quando comparados a outros cenários de variabilidade.

Na presença de autocorrelação espacial negativa, o desempenho dos modelos tende a ser melhor para a alta variabilidade existente nas regiões (Cenário 3). Em geral, os modelos tendem a apresentar melhores resultados para autocorrelações espaciais positivas no cenário de baixa concentração de pontos. Em um cenário de alta concentração de pontos, os melhores desempenhos são para valores de autocorrelação espacial negativa.

Para entender o comportamento dos dados intervalares no contexto de um modelo de regressão espacial, usamos o modelo CAR intervalar. Na Tabela 24, consideramos diferentes critérios para calcular as distâncias e diferentes parametrizações no cálculo para o desempenho do modelo. Um comportamento a ser observado nesse modelo é que os resultados não diferem quando comparado ao modelo de regressão linear intervalar. É como se a dependência espacial existente nos dados fosse captada pelo próprio intervalo. Com a finalidade de observar se os valores dos $\beta_i's$ impactam os resultados do modelo CAR, simulamos valores para esses $\beta_i's$.

Tabela 23 – Desempenho do modelo de regressão linear intervalar (MMREE) e o desvio padrão para diferentes configurações de dados intervalares.

ρ	**Cenário 1**		**Cenário 2**		**Cenário 3**	
	(Min;Max)	**(Cen;Amp)**	**(Min;Max)**	**(Cen;Amp)**	**(Min;Max)**	**(Cen;Amp)**
			1-**Vizinhos mais próximos**			
$-0,6$	0,0155(0,008)	0,0158(0,009)	0,1141(0,138)	0,1169(0,143)	0,0089(0,005)	0,0090(0,005)
$-0,5$	0,0181(0,010)	0,0185(0,010)	0,0417(0,034)	0,0423(0,035)	0,0084(0,005)	0,0085(0,005)
$-0,4$	0,0190(0,011)	0,0195(0,011)	0,0351(0,027)	0,0355(0,028)	0,0083(0,004)	0,0083(0,005)
$-0,3$	0,0187(0,010)	0,0192(0,011)	0,0367(0,029)	0,0372(0,029)	0,0083(0,004)	0,0084(0,005)
$-0,2$	0,0175(0,010)	0,0178(0,010)	0,0470(0,040)	0,0477(0,040)	0,0085(0,005)	0,0086(0,005)
$-0,1$	0,0154(0,008)	0,0158(0,009)	0,1022(0,117)	0,1040(0,119)	0,0089(0,005)	0,0089(0,005)
$0,1$	0,0112(0,006)	0,0114(0,006)	0,0353(0,024)	0,0357(0,024)	0,0110(0,007)	0,0111(0,007)
$0,2$	0,0092(0,005)	0,0094(0,005)	0,0183(0,012)	0,0185(0,012)	0,0144(0,010)	0,0146(0,010)
$0,3$	0,0075(0,004)	0,0078(0,004)	0,0116(0,007)	0,0117(0,007)	0,0292(0,023)	0,0297(0,024)
$0,4$	0,0061(0,003)	0,0061(0,003)	0,0079(0,004)	0,0080(0,004)	0,0447(0,033)	0,0454(0,033)
$0,5$	0,0046(0,002)	0,0047(0,002)	0,0055(0,0033)	0,0056(0,003)	0,0105(0,007)	0,0106(0,007)
$0,6$	0,0034(0,001)	0,0035(0,001)	0,0039(0,0022)	0,0039(0,002)	0,0053(0,003)	0,0053(0,003)
			2-**Vizinhos mais próximos**			
$-0,5$	0,5072(2,519)	0,5021(2,447)	0,0166(0,009)	0,0166(0,009)	0,0078(0,004)	0,0078(0,004)
$-0,4$	0,0779(0,050)	0,0793(0,050)	0,0153(0,008)	0,0153(0,008)	0,0076(0,004)	0,0075(0,004)
$-0,3$	0,9179(3,268)	0,9005(3,161)	0,0155(0,008)	0,0156(0,009)	0,0075(0,004)	0,0075(0,004)
$-0,2$	0,0380(0,022)	0,0385(0,022)	0,0182(0,011)	0,0183(0,011)	0,0076(0,004)	0,0076(0,004)
$-0,1$	0,0207(0,011)	0,0210(0,012)	0,0297(0,022)	0,0301(0,022)	0,0081(0,004)	0,0082(0,004)
0,1	0,0090(0,005)	0,0092(0,005)	0,0167(0,011)	0,0169(0,011)	0,0155(0,011)	0,0157(0,011)
0,2	0,0060(0,003)	0,0061(0,003)	0,0078(0,004)	0,0078(0,004)	0,0389(0,028)	0,0396(0,028)
0,3	0,0036(0,002)	0,0037(0,002)	0,0040(0,002)	0,0041(0,002)	0,0056(0,003)	0,0057(0,003)
0,4	0,0016(0,000)	0,0017(0,000)	0,0017(0,0009)	0,0017(0,000)	0,0018(0,001)	0,0018(0,001)
0,5	0,0003(0,000)	0,0003(0,000)	0,0003(0,000)	0,0003(0,000)	0,0003(0,000)	0,0003(0,000)
			Distância de Chebyshev			
$-0,6$	0,0909(0,061)	0,0916(0,061)	0,0156(0,008)	0,0157(0,008)	0,0077(0,004)	0,0077(0,004)
$-0,5$	0,2177(0,318)	0,2206(0,319)	0,0152(0,008)	0,0153(0,008)	0,0075(0,004)	0,0075(0,004)
$-0,4$	0,0713(0,045)	0,0722(0,044)	0,0163(0,009)	0,0164(0,009)	0,0075(0,004)	0,0076(0,004)
$-0,3$	0,0337(0,019)	0,0341(0,019)	0,0191(0,012)	0,0192(0,012)	0,0077(0,004)	0,0077(0,004)
$-0,2$	0,0227(0,013)	0,0230(0,013)	0,0258(0,018)	0,0261(0,018)	0,0080(0,004)	0,0081(0,004)
$-0,1$	0,0170(0,009)	0,0173(0,009)	0,0519(0,045)	0,0527(0,046)	0,0086(0,005)	0,0087(0,005)
0,1	0,0166(0,008)	0,0150(0,007)	0,0286(0,019)	0,0290(0,019)	0,0117(0,007)	0,0118(0,008)
0,2	0,0087(0,004)	0,0088(0,005)	0,0154(0,010)	0,0156(0,010)	0,0170(0,012)	0,0173(0,012)
0,3	0,0070(0,004)	0,0072(0,004)	0,0102(0,006)	0,0103(0,006)	0,0584(0,055)	0,0596(0,056)
0,4	0,0056(0,003)	0,0057(0,003)	0,0072(0,004)	0,0073(0,004)	0,0242(0,017)	0,0245(0,017)
0,5	0,0044(0,002)	0,0045(0,002)	0,0052(0,003)	0,0053(0,003)	0,0090(0,006)	0,0091(0,006)
0,6	0,0034(0,001)	0,0034(0,001)	0,0038(0,002)	0,0038(0,002)	0,0050(0,003)	0,0050(0,003)
			Critério da rainha			
$-0,6$	0,1065(0,077)	0,1078(0,074)	0,0159(0,009)	0,0160(0,009)	0,0075(0,004)	0,0075(0,004)
$-0,5$	0,0510(0,030)	0,0519(0,030)	0,0170(0,010)	0,0171(0,010)	0,0076(0,004)	0,0076(0,004)
$-0,4$	0,0344(0,020)	0,0350(0,020)	0,0189(0,012)	0,0191(0,012)	0,0077(0,004)	0,0077(0,004)
$-0,3$	0,0256(0,014)	0,0261(0,015)	0,0226(0,015)	0,0229(0,015)	0,0079(0,004)	0,0079(0,004)
$-0,2$	0,0202(0,012)	0,0205(0,012)	0,0312(0,023)	0,0316(0,024)	0,0082(0,005)	0,0083(0,005)
$-0,1$	0,0163(0,009)	0,0166(0,009)	0,0665(0,064)	0,0676(0,064)	0,0087(0,005)	0,0088(0,005)
0,1	0,0111(0,006)	0,0113(0,006)	0,0320(0,022)	0,0324(0,022)	0,0113(0,008)	0,0114(0,008)
0,2	0,0091(0,005)	0,0093(0,005)	0,0172(0,011)	0,0174(0,011)	0,0153(0,011)	0,0155(0,011)
0,3	0,0075(0,004)	0,0076(0,004)	0,0113(0,007)	0,0114(0,007)	0,0329(0,028)	0,0334(0,028)
0,4	0,0061(0,003)	0,0062(0,003)	0,0079(0,005)	0,0080(0,005)	0,0439(0,033)	0,0446(0,035)
0,5	0,0048(0,003)	0,0049(0,003)	0,0057(0,003)	0,0058(0,004)	0,0112(0,008)	0,0113(0,008)
0,6	0,0037(0,002)	0,0037(0,002)	0,0041(0,005)	0,0041(0,005)	0,0057(0,004)	0,0058(0,004)

Fonte: Elaborada pela autora.

Tabela 24 – Desempenho do modelo CAR intervalar (MMREE) e o desvio padrão para diferentes configurações de dados intervalares.

ρ	**Cenário 1**		**Cenário 2**		**Cenário 3**	
	(Min;Max)	**(Cen;Amp)**	**(Min;Max)**	**(Cen;Amp)**	**(Min;Max)**	**(Cen;Amp)**
			1-**Vizinho mais próximo**			
$-0,6$	0,0154(0,008)	0,0158(0,008)	0,1144(0,139)	0,1171(0,144)	0,0039(0,002)	0,0039(0,002)
$-0,5$	0,0181(0,010)	0,0185(0,010)	0,0418(0,034)	0,0424(0,035)	0,0085(0,005)	0,0085(0,005)
$-0,4$	0,0191(0,011)	0,0196(0,011)	0,0352(0,027)	0,0356(0,028)	0,0083(0,000)	0,0084(0,005)
$-0,3$	0,0188(0,010)	0,0192(0,011)	0,0368(0,029)	0,0373(0,029)	0,0084(0,005)	0,0084(0,005)
$-0,2$	0,0174(0,010)	0,0178(0,010)	0,0471(0,040)	0,0478(0,041)	0,0085(0,005)	0,0086(0,005)
$-0,1$	0,0155(0,008)	0,0158(0,009)	0,1025(0,117)	0,1042(0,119)	0,0089(0,005)	0,0090(0,005)
0,1	0,0028(0,005)	0,0028(0,006)	0,0353(0,024)	0,0358(0,025)	0,0110(0,007)	0,0112(0,007)
0,2	0,0094(0,005)	0,0095(0,005)	0,0183(0,012)	0,0186(0,012)	0,0145(0,010)	0,0147(0,010)
0,3	0,0076(0,004)	0,0078(0,004)	0,0116(0,007)	0,0118(0,007)	0,0112(0,021)	0,0113(0,022)
0,4	0,0061(0,003)	0,0062(0,003)	0,0080(0,004)	0,0080(0,005)	0,0448(0,033)	0,0455(0,033)
0,5	0,0047(0,002)	0,0048(0,002)	0,0056(0,003)	0,0056(0,003)	0,0105(0,007)	0,0107(0,007)
0,6	0,0035(0,002)	0,0036(0,000)	0,0039(0,002)	0,0039(0,002)	0,0053(0,003)	0,0054(0,003)
			2-**Vizinhos mais próximos**			
$-0,5$	0,5068(2,517)	0,5024(2,44)	0,0166(0,009)	0,0166(0,009)	0,0078(0,004)	0,0078(0,004)
$-0,4$	0,0779(0,050)	0,0794(0,051)	0,0153(0,008)	0,0153(0,008)	0,0076(0,004)	0,0076(0,004)
$-0,3$	0,9182(3,269)	0,8983(3,132)	0,0156(0,008)	0,0157(0,009)	0,0075(0,004)	0,0075(0,004)
$-0,2$	0,0380(0,022)	0,0386(0,022)	0,0182(0,011)	0,0182(0,011)	0,0077(0,004)	0,0077(0,004)
$-0,1$	0,0207(0,011)	0,0210(0,012)	0,0298(0,022)	0,0298(0,022)	0,0082(0,004)	0,0082(0,004)
0,1	0,0091(0,005)	0,0092(0,005)	0,0168(0,011)	0,0169(0,011)	0,0156(0,011)	0,0157(0,011)
0,2	0,0060(0,003)	0,0061(0,003)	0,0078(0,004)	0,0079(0,004)	0,0390(0,028)	0,0395(0,028)
0,3	0,0037(0,002)	0,0037(0,002)	0,0041(0,002)	0,0041(0,002)	0,0056(0,003)	0,0057(0,003)
0,4	0,0017(0,001)	0,0017(0,001)	0,0017(0,001)	0,0017(0,001)	0,0019(0,001)	0,0019(0,001)
0,5	0,0003(0,000)	0,0003(0,000)	0,0003(0,000)	0,0003(0,000)	0,0002(0,000)	0,0002(0,000)
			Distância de Chebyshev			
$-0,6$	0,0909(0,061)	0,0916(0,061)	0,0156(0,008)	0,0157(0,008)	0,0076(0,006)	0,0074(0,006)
$-0,5$	0,2176(0,318)	0,2207(0,319)	0,0035(0,007)	0,0036(0,007)	0,0075(0,004)	0,0075(0,004)
$-0,4$	0,0213(0,039)	0,0216(0,039)	0,0163(0,009)	0,0164(0,009)	0,0053(0,005)	0,0053(0,005)
$-0,3$	0,0337(0,019)	0,0342(0,020)	0,0191(0,012)	0,0192(0,012)	0,0077(0,004)	0,0077(0,004)
$-0,2$	0,0228(0,013)	0,0231(0,013)	0,0042(0,013)	0,0043(0,013)	0,0018(0,004)	0,0018(0,004)
$-0,1$	0,0171(0,009)	0,0173(0,009)	0,0519(0,045)	0,0527(0,046)	0,0084(0,005)	0,0085(0,005)
0,1	0,0108(0,006)	0,0110(0,006)	0,0286(0,019)	0,0290(0,019)	0,0045(0,007)	0,0047(0,008)
0,2	0,0087(0,004)	0,0089(0,005)	0,0154(0,010)	0,0156(0,010)	0,0170(0,012)	0,0176(0,013)
0,3	0,0071(0,004)	0,0072(0,004)	0,0102(0,006)	0,0103(0,006)	0,0060(0,035)	0,0069(0,038)
0,4	0,0057(0,003)	0,0058(0,003)	0,0072(0,004)	0,0073(0,004)	0,0048(0,013)	0,0052(0,014)
0,5	0,0045(0,002)	0,0046(0,002)	0,0052(0,003)	0,0053(0,003)	0,0010(0,004)	0,0011(0,004)
0,6	0,0009(0,001)	0,0009(0,001)	0,0009(0,001)	0,0009(0,001)	0,0050(0,003)	0,0050(0,003)
			Critério da rainha			
$-0,6$	0,1065(0,077)	0,1080(0,075)	0,0157(0,009)	0,0160(0,009)	0,0012(0,003)	0,0012(0,003)
$-0,5$	0,0509(0,030)	0,0519(0,031)	0,0167(0,010)	0,0171(0,010)	0,0076(0,004)	0,0076(0,004)
$-0,4$	0,0343(0,020)	0,0350(0,020)	0,0186(0,012)	0,0190(0,012)	0,0077(0,004)	0,0077(0,004)
$-0,3$	0,0256(0,014)	0,0260(0,015)	0,0224(0,015)	0,0228(0,015)	0,0079(0,004)	0,0079(0,004)
$-0,2$	0,0202(0,011)	0,0205(0,011)	0,0045(0,016)	0,0047(0,016)	0,0013(0,003)	0,0013(0,003)
$-0,1$	0,0163(0,009)	0,0165(0,009)	0,0663(0,063)	0,0674(0,064)	0,0088(0,005)	0,0088(0,005)
0,1	0,0110(0,006)	0,0112(0,006)	0,0320(0,022)	0,0323(0,022)	0,0113(0,007)	0,0114(0,007)
0,2	0,0091(0,005)	0,0093(0,005)	0,0172(0,011)	0,0173(0,011)	0,0153(0,011)	0,0154(0,011)
0,3	0,0075(0,004)	0,0076(0,004)	0,0113(0,007)	0,0113(0,007)	0,0329(0,027)	0,0334(0,027)
0,4	0,0061(0,003)	0,0062(0,003)	0,0080(0,004)	0,0080(0,005)	0,0319(0,036)	0,0319(0,036)
0,5	0,0048(0,002)	0,0049(0,002)	0,0057(0,003)	0,0058(0,003)	0,0112(0,007)	0,0113(0,007)
0,6	0,0037(0,002)	0,0037(0,002)	0,0041(0,002)	0,0041(0,002)	0,0057(0,003)	0,0058(0,003)

Fonte: Elaborada pela autora.

Tabela 25 – Performance para o modelo de regressão linear CAR intervalar (MMREE) para diferentes configurações de dados, no qual a região de vizinhança é identificada usando uma distância clássica.

ρ	**Cenário 1**		**Cenário 2**		**Cenário 3**	
	(Min;Max)	**(Cen;Amp)**	**(Min;Max)**	**(Cen;Amp)**	**(Min;Max)**	**(Cen;Amp)**
	1-Vizinho mais próximo					
-0,6	0,8628	0,8867	0,5732	0,5038	0,5013	0,5068
-0,5	0,7735	0,7681	0,4174	0,4178	0,3571	0,3735
-0,4	0,7695	0,7575	0,3898	0,3849	0,3825	0,4295
-0,3	0,7660	0,7680	1,0430	0,8818	0,3937	0,4139
-0,2	0,7793	0,7606	0,4126	0,4175	0,5207	0,5174
-0,1	0,6269	0,5720	0,4796	0,5783	0,368	0,3686
0,1	0,5455	0,5169	0,3733	0,3809	0,3794	0,3934
0,2	0,7006	0,6376	0,4182	0,4328	0,3545	0,3772
0,3	0,6732	0,5856	0,4122	0,4059	0,3473	0,3719
0,4	0,5932	0,5025	0,4297	0,4276	0,4508	0,6043
0,5	0,3917	0,3676	0,3197	0,3162	0,4131	0,4674
0,6	0,3202	0,3797	0,4047	0,4070	0,3450	0,3675
	2-Vizinhos mais próximos					
-0,6	0,2693	0,2705	0,3759	0,3550	0,3800	0,3823
-0,5	0,7443	0,7796	0,4476	0,4591	0,4922	0,5999
-0,3	0,5851	0,6138	0,6879	0,7089	0,7102	0,7556
-0,2	0,6306	0,6389	0,4464	0,4591	0,5972	0,6114
-0,1	0,7682	0,8302	0,4476	0,4394	0,6921	0,6381
0,1	0,4693	0,4371	0,4770	0,4745	0,5425	0,5269
0,2	0,4148	0,4136	0,5199	0,504	0,4607	0,4600
0,3	0,2804	0,2804	0,4193	0,4112	1,8840	1,4225
0,4	0,2564	0,2569	0,3759	0,3550	0,3800	0,3823
0,5	0,2564	0,2569	0,3759	0,3550	0,3800	0,3823
0,6	0,2564	0,2569	0,3759	0,3550	0,3800	0,3823
	Critério da rainha					
-0,6	0,9130	1,0410	0,4208	0,4282	0,3920	0,4111
-0,5	0,7429	0,7367	0,3689	0,3763	0,4178	0,4196
-0,4	1,9405	2,2566	0,3565	0,3786	0,3626	0,4472
-0,3	1,7517	1,9802	0,4344	0,4194	0,3885	0,412
-0,2	0,5087	0,5063	0,3350	0,3498	0,6322	0,6055
-0,1	0,6148	0,6271	0,3760	0,3799	0,3916	0,4134
0,1	0,7492	0,6461	0,3565	0,3635	0,3501	0,3585
0,2	0,5194	0,4981	0,4348	0,3914	0,3535	0,3788
0,3	0,3791	0,3731	0,3746	0,3809	0,6241	0,4774
0,4	0,3927	0,3626	0,4205	0,4783	0,3462	0,3866
0,5	0,4844	0,4973	0,3499	0,3415	0,3289	0,3568
0,6	0,2845	0,2849	0,3687	0,3657	0,3229	0,3225
	Distância de Chebyshev					
-0,6	0,6544	0,6040	0,3914	0,4114	0,5077	0,5353
-0,5	0,7087	0,6472	0,4887	0,5433	0,4538	0,5493
-0,4	0,7660	0,6689	0,3792	0,4290	0,4036	0,4540
-0,3	1,4276	1,2697	0,4699	0,5253	0,4130	0,4587
-0,2	0,6777	0,5285	0,3753	0,3905	0,6231	0,6438
-0,1	1,6952	0,6648	0,7344	0,6956	0,4156	0,4459
0,1	1,2263	0,9932	0,4152	0,4089	0,5344	0,7552
0,2	0,7257	0,4582	0,8143	0,9129	0,4512	0,4499
0,3	0,7956	0,5738	0,5689	0,5406	0,3835	0,3959
0,4	0,7975	0,4396	0,3990	0,4188	0,4679	0,5433
0,5	0,6225	0,3409	0,4370	0,3975	1,1550	0,7588
0,6	0,3372	0,2993	0,5299	0,5548	0,4716	0,4120

Fonte: Elaborada pela autora.

Tabela 26 – Performance para o modelo de regressão linear CAR intervalar (MMREE) para diferentes configurações de dados, no qual a região de vizinhança é identificada usando uma distância intervalar.

ρ	**Cenário 1**		**Cenário 2**		**Cenário 3**	
	(Min; Max)	**(Cen; Amp)**	**(Min; Max)**	**(Cen; Amp)**	**(Min; Max)**	**(Cen; Amp)**
	Distância euclideana intervalar					
$-0,6$	0,6519	0,6080	0,7884	0,8011	1,0084	0,6622
$-0,5$	0,6945	0,7063	0,4739	0,4881	0,4683	0,4536
$-0,4$	0,8373	0,7431	1,0910	1,1082	0,456	0,4783
$-0,3$	0,9599	1,0580	0,5221	0,5571	0,5234	0,5265
$-0,2$	0,6621	0,6153	0,4468	0,4617	0,4737	0,5181
$-0,1$	0,7556	0,7659	0,5195	0,5537	0,4642	0,4841
0,1	0,5831	0,5106	0,4356	0,4368	0,4674	0,4752
0,2	0,5369	0,4869	0,4807	0,5251	0,4554	0,4897
0,3	0,5171	0,4890	0,4755	0,4627	0,5325	0,5243
0,4	0,4611	0,4299	0,4763	0,4912	0,6178	0,5571
0,5	0,3974	0,3813	0,3957	0,4090	0,4567	0,4649
0,6	0,3239	0,3350	0,3960	0,3865	0,4451	0,4776
	Distância de city-block intervalar					
$-0,6$	0,7065	0,6417	0,8951	0,7887	1,3036	0,5907
$-0,5$	0,7533	0,7242	0,5040	0,4412	0,4620	0,4181
$-0,4$	0,8572	0,7418	1,3736	1,1182	0,4719	0,4433
$-0,3$	0,7455	0,7302	0,5766	0,5284	0,5832	0,4866
$-0,2$	0,7928	0,7729	0,4754	0,4342	0,5050	0,4817
$-0,1$	0,7855	0,8116	0,5618	0,4985	0,4604	0,4162
0,1	0,5628	0,5207	0,4505	0,4162	0,5155	0,4471
0,2	0,4828	0,4577	0,5536	0,4958	0,5089	0,4794
0,3	0,4354	0,4139	0,4782	0,4379	0,5315	0,4287
0,4	0,4838	0,4611	0,5535	0,4396	0,5749	0,4801
0,5	0,3738	0,3614	0,4009	0,3796	0,4709	0,4113
0,6	0,3163	0,3298	0,4808	0,4426	0,4793	0,4433
	Distância de Hausdorff intervalar					
$-0,6$	1,0797	0,9224	1,0288	0,8188	1,4491	0,6620
$-0,5$	0,8609	0,7982	0,5916	0,4682	0,5201	0,4306
$-0,4$	0,9066	1,0611	0,6038	0,5227	0,5102	0,4495
$-0,3$	0,7467	0,6583	0,5803	0,5173	0,5838	0,4842
$-0,2$	0,8858	0,8483	0,4757	0,4300	0,5020	0,4530
$-0,1$	0,7790	0,8035	0,5910	0,5168	0,4861	0,4235
0,1	0,6113	0,5175	0,4620	0,4049	0,516	0,4477
0,2	0,4965	0,4562	0,5435	0,4477	0,5072	0,4512
0,3	0,4574	0,3961	0,5139	0,4206	0,6006	0,4538
0,4	0,5254	0,4663	0,8615	0,4596	0,6756	0,4839
0,5	0,3898	0,3516	0,4314	0,3859	0,6250	0,5354
0,6	0,3156	0,3144	0,5144	0,4098	0,7170	0,4604

Fonte: Elaborada pela autora.

As Tabelas 25 e 26 apresentam a performance para o modelo CAR intervalar considerando o Algoritmo 3, no qual os valores simulados para β_0 e β_1 são gerados de forma aleatória. Nessas

configurações, β_0 e β_1 seguem distribuição uniforme no intervalo [1,5] e [1,3], respectivamente. Além disso, consideramos diferentes critérios de vizinhança. Na Tabela 25, consideramos os critérios knn, contiguidade da rainha e distância de Chebyshev para geração da região de vizinhança. Na Tabela 26, consideramos distâncias intervalares para o cálculo da matriz de vizinhança.

A partir da Tabela 25, é possível analisar o desempenho do modelo levando em consideração diferentes cenários de variabilidade e parametrizações. Podemos observar que diferentes distâncias resultam em diferentes valores para a performance do modelo. Dentre as métricas observadas, o 2-vizinhos mais próximos resultou em um melhor desempenho. Ao observarmos os diferentes valores de correlação ρ, nota-se um melhor desempenho do modelo para dados com uma correlação positiva. De maneira geral, para esses resultados a parametrização mínimo e máximo tende a apresentar um melhor desempenho no cenário de menor variabilidade nos dados (Cenário 1). Por outro lado, quando a variabilidade aumenta (Cenário 3), a parametrização centro e amplitude tende a aprresentar um melhor comportamento.

Ao analisarmos os resultados do modelo CAR intervalar utilizando distâncias intervalares (Tabela 26), podemos verificar o desempenho do modelo com diferentes cenários, parametrizações e distintos valores de ρ. Inicialmente, podemos ver que esse modelo resulta em um melhor desempenho para dados com correlação positiva. Além disso, observa-se que o cenário de variabilidade interfere no desempenho do modelo. Se fixarmos os cenários 1 e 2, no qual a região de vizinhaça é identificada usando a distância euclideana intervalar, podemos observar que o desempenho do modelo tende a melhorar para valores negativos de ρ e maior variabilidade (cenário 3). Esse comportamento difere se consideramos valores negativos e uma maior variabilidade. Comportamento similar também é observado para outras distâncias utilizadas. Dessa forma, o desempenho desse modelo é afetado pelo tipo de correlação presente nos dados e a variabilidade existente. Com respeito às parametrizações abordadas, a partir dessa tabela, é possível verificar que o desempenho do modelo tende a apresentar resultados semelhenates. No entanto, a parametrização centro e amplitude tende a apresentar o melhor desempenho. Para isso, considere a distância de city-block intervalar. Nesse cenário, o desempenho da parametrização centro e amplitude apresentou um melhor desempenho, independentemente da variabilidade existente ou da correlação presente nos dados.

Ao compraramos os resultados das Tabelas 25 e 26, em que a diferença entre elas consiste no tipo de distância utilizada para a identificação da região de vizinhança, observamos que o modelo CAR intervalar tende a apresentar melhores resultados quando utilizamos o 2- vizinhos

mais próximos para identificar a região de vizinhança.

4.4 APLICAÇÕES COM CONJUNTOS DE DADOS AGREGADOS

Nesta seção, utilizamos dados reais georreferenciados para examinar as aplicações dos modelos propostos neste estudo. Para a análise da autocorrelação espacial, consideramos o índice de Moran (MORAN, 1950). Dois conjuntos clássicos foram agregados para formar um conjunto de dados simbólicos. A primeira aplicação refere-se aos dados de mortalidade infantil para o estado de Pernambuco e a segunda aplicação utiliza informações sobre a renda per capita no nordeste brasileiro.

4.4.1 Aplicação I: Dados de mortalidade infantil

Consideremos um conjunto de dados clássicos agregados para gerar uma variável simbólica do tipo intervalo. Este banco de dados também contém informações sobre a localização geográfica. Assim, estamos interessados na análise exploratória de dados de contagem que são espacialmente 'rotulados' (ou seja, são indexados por localização geográfica).

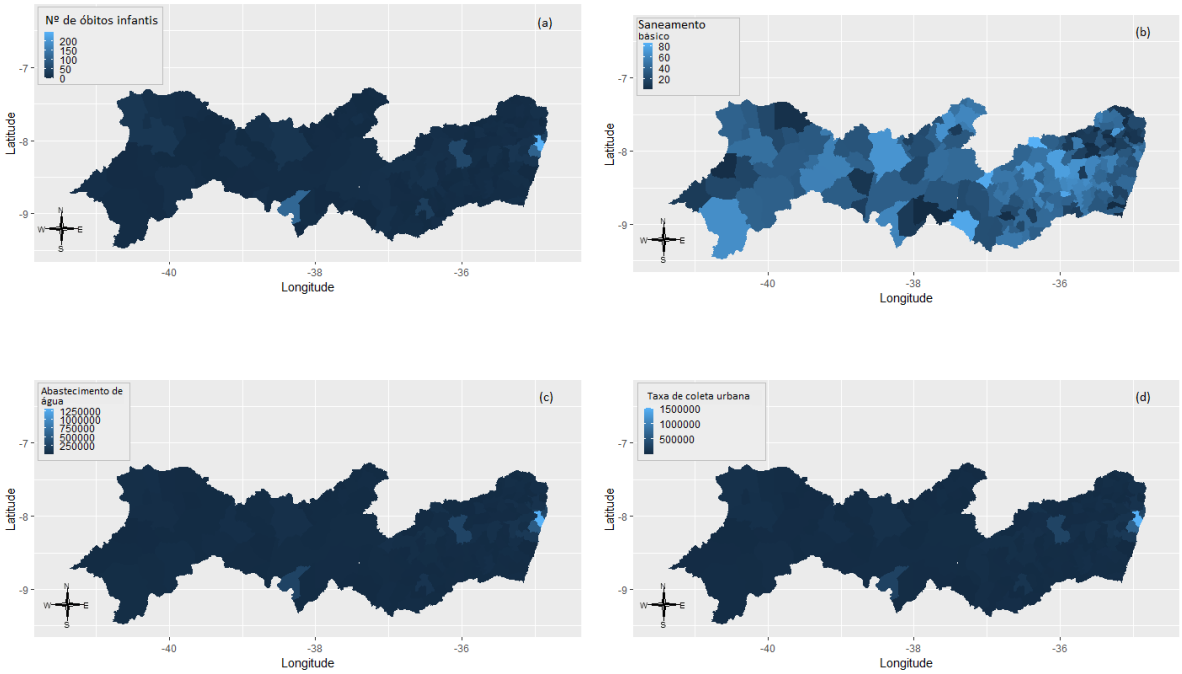
Aplicamos o modelo de regressão linear intervalar da literatura de dados simbólicos e o modelo proposto para dados intervalares no contexto de dados autocorrelacionados. Os dados utilizados para esta análise foram obtidos no site Datasus, do governo federal, que disponibiliza informações públicas sobre o Serviço Único de Saúde (SUS) no país.

Esses dados referem-se ao estado de Pernambuco, localizado na região Nordeste do Brasil. As variáveis analisadas são o número de óbitos infantis, taxa de coleta de lixo urbano, taxa de abastecimento de água e taxa de saneamento básico. De acordo com a Figura 19, podemos visualizar o comportamento espacial dessas variáveis.

Utilizamos o índice de Moran global para analisar a autocorrelação espacial entre as áreas (municípios). O índice de Moran fornece uma maneira de obter a autocorrelação espacial global de cada variável. Assim, é possível detectar a presença de dados autocorrelacionados.

A Tabela 27 apresenta os índices de autocorrelação espacial de cada variável em estudo e seus respectivos níveis de significância baseados no teste de autocorrelaçao espacial de Moran. Para o processo de geração desses dados, utilizamos o critério de 1-vizinho mais próximo para identificar a matriz da região de vizinhança. Notamos que os níveis descritivos (p-valor) são inferiores a 0,05 (nível de significância), concluindo que todas as variáveis possuem

Figura 19 – Distribuição espacial dos municípios considerando o número de óbitos infantis (a), saneamento básico (b), abastecimento de água (c) e coleta de lixo urbano (d).



Fonte: Elaborada pela autora.

autocorrelação espacial significativa a 5% de probabilidade.

Tabela 27 – Índices de autocorrelação espacial para dados clássicos.

Variável	I	Esperança	Variância	P-valor
Número de óbitos infantis	0,2690	-0,0053	0,0053	0,0001
Coleta de lixo urbano	0,3528	-0,0054	0,0042	0,0000
Abastecimento de água	0,3908	-0,0054	0,0040	0,0000
Saneamento básico	0,2749	-0,0054	0,0089	0,0030

Fonte: Elaborada pela autora.

A agregação em classes e sua descrição por dados simbólicos são úteis para extrair conhecimento de grandes bancos de dados. Nesse contexto, agregamos informações de primeiro nível (cidades) em um segundo nível que representa os grupos (conjunto de cidades). Dessa forma, podemos descobrir novos tipos de informações para uma determinada região que não estão disponíveis no nível individual.

O conjunto de dados clássico refere-se a informações sobre as 185 cidades do Estado de Pernambuco. Agregamos esses dados em 19 microrregiões. Em seguida, obtemos os valores mínimo e máximo de cada variável. A Tabela 28 mostra uma pequena parte do conjunto de dados de intervalo. Baseado na técnica de seleção de variáveis, stepwise, selecionamos três

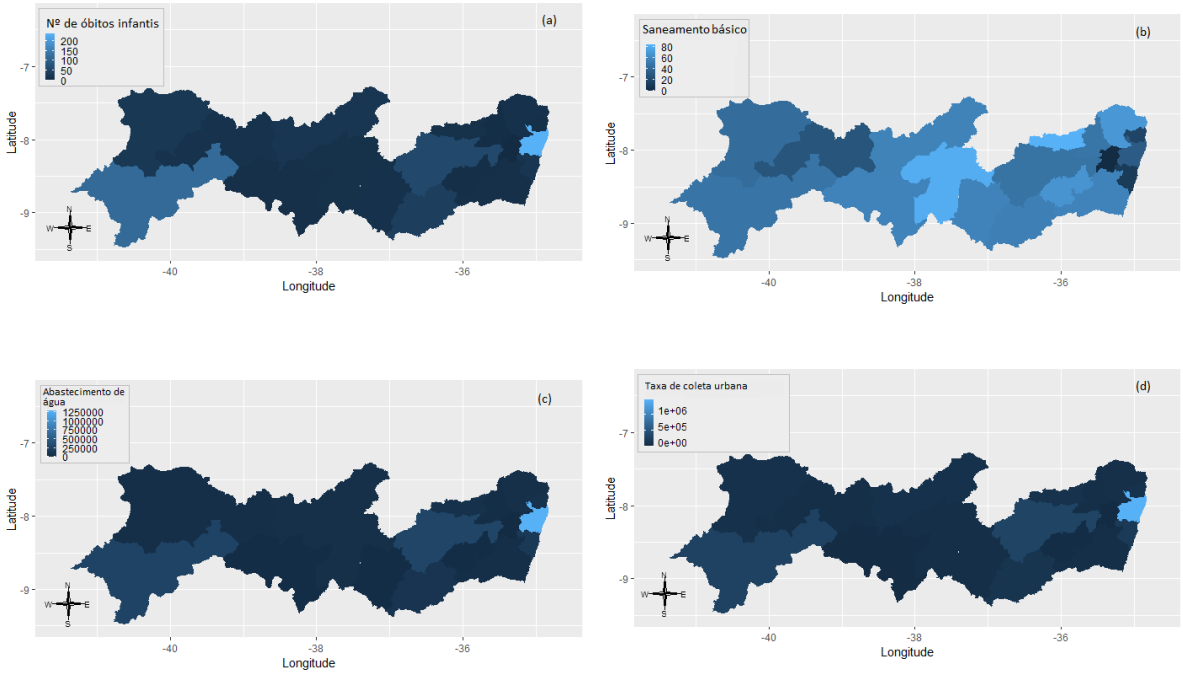
variáveis para prever a variável resposta. Aqui, para a variável dependente (número de óbitos infantis), consideramos três variáveis intervalares independentes: taxa de saneamento básico, taxa de abastecimento de água e taxa de coleta de lixo urbano.

Tabela 28 – Conjunto de dados simbólico intervalar para o estado de Pernambuco.

Região	**Centro da latitude**	**Centro da longitude**	**Número de óbitos infantis**	$\cdots$	**Coleta de lixo urbano**
R_1	-8,100	-35,331	[1; 58]	$\cdots$	[28,24; 74,92]
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	
R_{19}	-7,874	-35,935	[0; 11]	$\cdots$	[2,49; 85,54]

Fonte: Elaborada pela autora.

Figura 20 – Distribuição espacial por microrregiões para as variáveis número de óbitos infantis (a), saneamento básico (b), abastecimento de água (c) e coleta de lixo urbano (d).



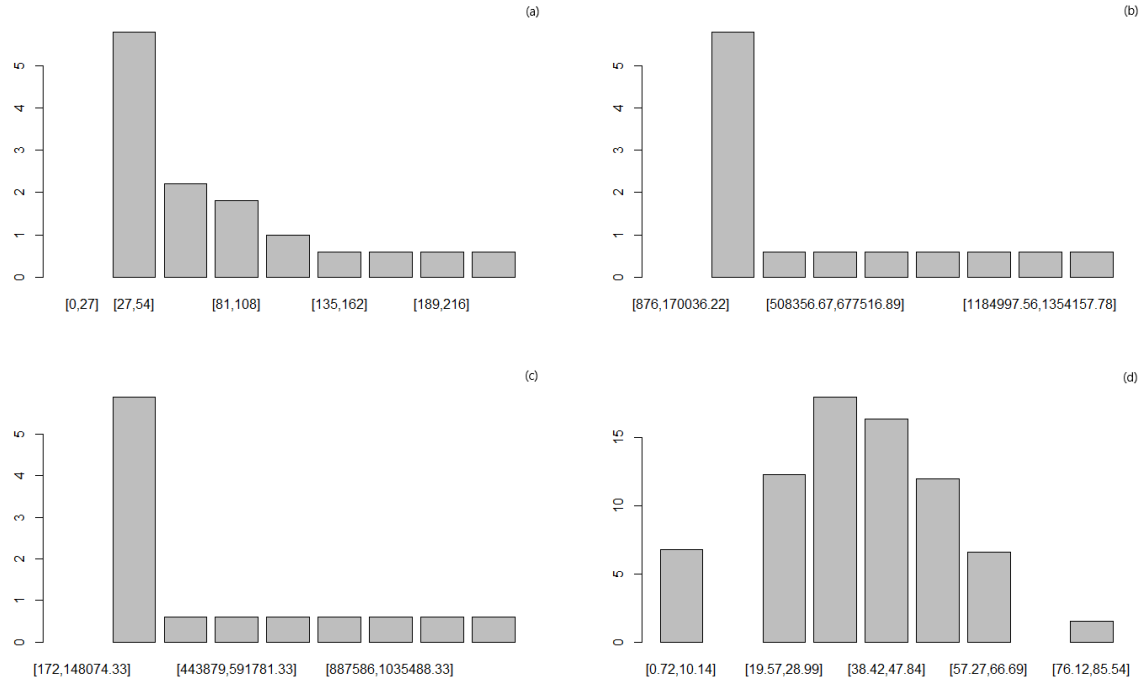
Fonte: Elaborada pela autora.

A Figura 20 mostra a visualização das variáveis intervalares para as microrregiões de Pernambuco, onde observamos a distribuição espacial das regiões. Na Figura 20 (a), podemos observar a distribuição espacial do número de óbitos infantis indicando um comportamento espacial heterogêneo. Os pontos mais claros no mapa são as regiões com alta variabilidade no número de óbitos. De acordo com a Figura 20 (b), podemos observar uma alta variabilidade entre as microrregiões, indicando um comportamento heterogêneo dentro das regiões ao analisarmos a variável saneamento básico. A Figura 20 (c) identifica a microrregião do Recife

como a área de maior variabilidade para a taxa de abastecimento de água. Na Figura 20 (d), observamos que algumas microrregiões apresentam alta variabilidade ao analisar a coleta de lixo urbana. Assim, este mapa permite observar uma variabilidade substancial nas variáveis em estudo quanto ao seu caráter espacial.

Agora, podemos apresentar a representação gráfica da distribuição das variáveis intervalares através de histogramas. Consideremos a Figura 21. Podemos notar que as variáveis possuem comportamentos diferentes com forte assimetria positiva. Por exemplo, na Figura 21(a), a partir da distribuição dos intervalos para o número de óbitos infantis, verifica-se que o primeiro intervalo apresenta a frequência máxima.

Figura 21 – Histogramas para variáveis intervalares: número de óbitos infantis (a); coleta de lixo urbano (b); taxa de abastecimento de água (c); taxa de saneamento básico (d), referente ao Estado de Pernambuco.



Fonte: Elaborada pela autora.

Para complementar a abordagem proposta, podemos empregar modelos de regressão para dados intervalares da literatura de SDA. Além disso, apresentamos resultados obtidos com o modelo de regressão espacial para dados intervalares. Mais precisamente, aplicamos modelos de regressão linear intervalar e o modelo SDA para intervalos a um conjunto de dados simbólico. A avaliação de desempenho desses modelos de regressão é baseada na MMREE, que é dada pela Equação 4.9, sendo calculada para cada modelo de regressão com diferentes parametrizações. A Tabela 29 mostra os desempenhos correspondentes para esses modelos.

Tabela 29 – Avaliação de desempenho dos modelos para o estado de Pernambuco, com base na MMREE.

Modelos	**(Min; Max)**	**(Cen; Amp)**
Regressão linear intervalar	1,3662	2,2620
Regressão CAR intervalar	0,9199	1,9518

Fonte: Elaborada pela autora.

De acordo com a Tabela 29, podemos verificar o comportamento de ambos os modelos sob diferentes parametrizações. É possível observar que o modelo CAR intervalar utilizando a parametrização mínimo e máximo apresentou melhor desempenho. Observamos que o modelo que considera as informações espaciais e a variabilidade existente oferece melhor desempenho na comparação dos resultados obtidos. Nesse contexto, a localização dos fenômenos analisados é de suma importância para obter um melhor desempenho.

4.4.2 Aplicação II: Renda Per Capita

A renda per capita é um índice socioeconômico relevante, fornecendo uma medida do desenvolvimento econômico de uma região de interesse. Seu cálculo é obtido dividindo-se o Produto Interno Bruto (PIB) pelo número total de habitantes. Segundo dados do Instituto Brasileiro de Geografia e Estatística (IBGE), a renda per capita do Nordeste é de R$ 1,588. Com esse desempenho, a região Nordeste apresenta um valor inferior às demais regiões do Brasil, como a região Norte (R$1,687); Sudeste (R$ 2,650); Centro-Oeste (R$ 2,506) e Sul com R$ 2,549.

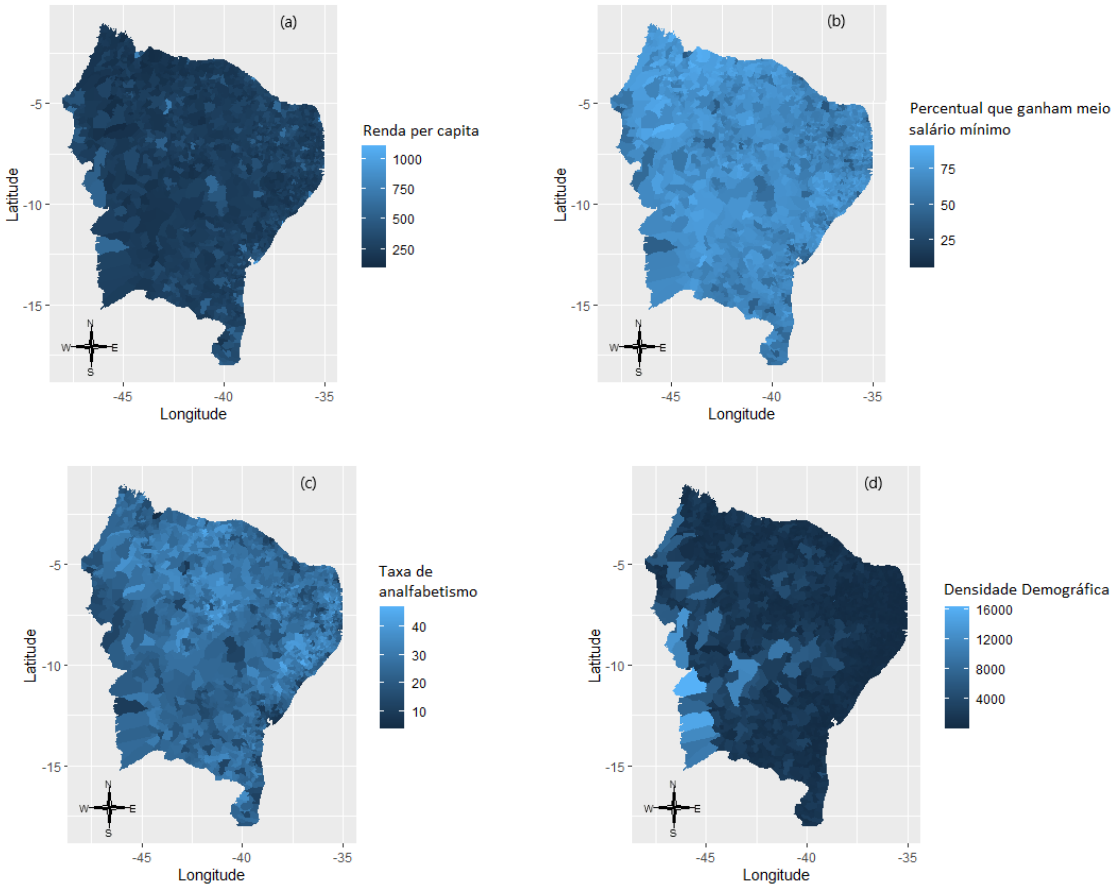
Para examinar os dados referentes às variáveis socioeconômicas da região Nordeste do Brasil, utilizamos metodologias espaciais. Buscamos compreender o comportamento espacial da renda per capita considerando-a como uma variável simbólica intervalar. Obtivemos a base cartográfica do Nordeste brasileiro, disponível no site do IBGE. Essa base contém informações sobre os 1794 municípios da região Nordeste. O processo de agregação dos municípios resultou em 188 microrregiões. Em seguida, utilizamos ferramentas de análise espacial e análise de dados simbólicos para caracterizar algumas variáveis socioeconômicas nessas 188 microrregiões.

A Figura 22 apresenta a distribuição espacial das seguintes variáveis: renda per capita, percentual das famílias que recebem meio salário mínimo, taxa de analfabetismo e densidade demográfica.

Semelhante à aplicação anterior, utilizamos o teste de Moran global para observar o nível de

autocorrelação espacial entre as áreas (municípios). De acordo com os resultados da Tabela 30, observamos que as variáveis percentual de famílias que ganham meio salário mínimo e taxa de analfabetismo possuem forte autocorrelação espacial. Podemos entender isso através da semelhança entre populações. Mais precisamente, municípios com alto percentual de famílias que ganham meio salário mínimo ou taxa de analfabetismo tendem a ser vizinhas de populações com características semelhantes.

Figura 22 – Distribuição espacial dos municípios considerando as variáveis renda per capita (a), percentual das famílias que ganham meio salário mínimo (b), taxa de analfabetismo (c) e densidade demográfica (d).



Fonte: Elaborada pela autora.

Tabela 30 – Índices de autocorrelação espacial para a região Nordeste do Brasil.

Variável	I	Esperança	Variância	P-Valor
Renda per capita	0,3204	-0,0005	0,0008	0,0000
Perc. meio salário mínimo	0,4740	-0,0005	0,0009	0,0000
Taxa de analfabetismo	0,5886	-0,0005	0,0009	0,0000
Densidade demográfica	0,3144	-0,0005	0,0008	0,0000

Fonte: Elaborada pela autora.

Para construir as variáveis simbólicas do tipo intervalo, agregamos os municípios por microrregiões considerando os valores máximo e mínimo para cada região (Tabela 31). A Figura 23 refere-se ao comportamento da variabilidade dentro de cada região, mostrando os intervalos correspondentes para essas regiões. Embora cada variável apresente um comportamento diferente, é possível identificar microrregiões críticas com alta variabilidade em relação ao fenômeno estudado.

A Figura 24 mostra a representação gráfica da distribuição das variáveis intervalares por meio de histogramas. Observamos na Figura 24 (a) e 24 (d), que as variáveis possuem o maior pico do histograma para sua primeira faixa de valores, resultando em um comportamento assimétrico. Notamos que os pequenos intervalos apresentam altas frequências, enquanto os grandes têm uma frequência modesta. A variável percentual de famílias que ganham meio salário mínimo tem comportamento assimétrico negativo, indicando que a maioria de seus intervalos são mais frequentes para os maiores valores de máximos e mínimos (Figura 24(b)). Observamos um comportamento distinto na variável taxa de analfabetismo, em que há uma simetria na distribuição dos intervalos (Figura 24(c)).

Tabela 31 – Conjunto de dados simbólicos intervalares para a região Nordeste do Brasil.

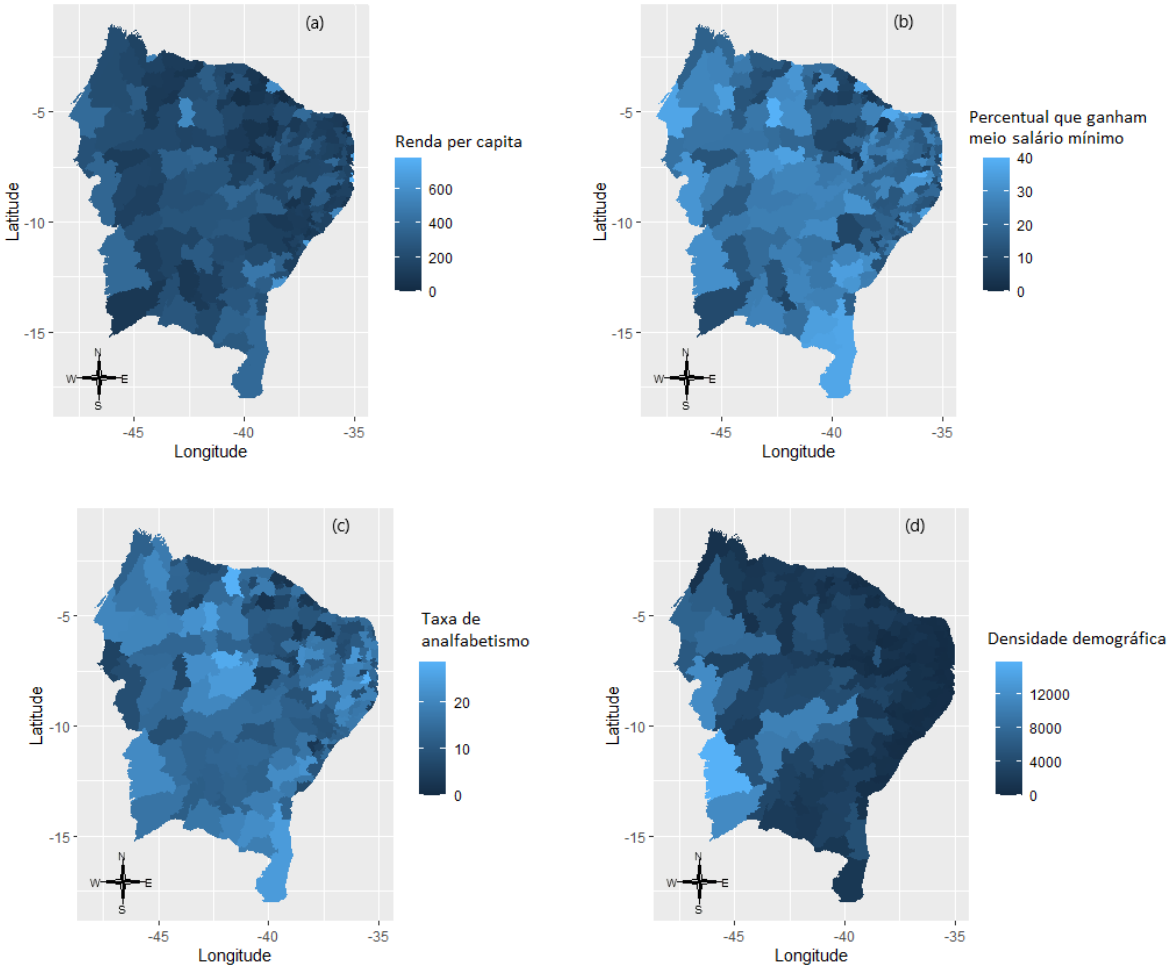
Região	**Centro da Latitude**	**Centro da Longitude**	**Renda per capita**	$\cdots$ $\cdots$	**Densidade demográfica**
R_1	-2,16	-44,77	[120,33; 272,32]	$\cdots$	[213; 1486,7]
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
R_{188}	-23,87	-47,58	[148,67; 309,25]	$\cdots$	[244,8; 6590,5]

Fonte: Elaborada pela autora.

Novamente, usamos modelos de regressão para nos ajudar a entender a abordagem. A Tabela 32 apresenta os desempenhos desses modelos de regressão.

Agora, vamos comentar brevemente sobre o uso de um método de regressão linear intervalar e o modelo CAR intervalar. Verificamos que os modelos espaciais e o modelo CAR intervalar tendem a capturar a dependência espacial existente nos dados. A Tabela 32 apresenta os desempenhos dessas abordagens. De acordo com os resultados obtidos, ambos os modelos têm um desempenho semelhante. O modelo de regressão linear para intervalos revela a variabilidade nos dados e a dependência espacial existente nos dados, pois apresenta resultados semelhantes a um modelo de regressão para dados espaciais.

Figura 23 – Distribuição espacial para as micorregiões das variáveis renda per capita (a), percentual de famílias que ganham meio salário mínimo (b), taxa de analfabetismo (c) e densidade demográfica (d).



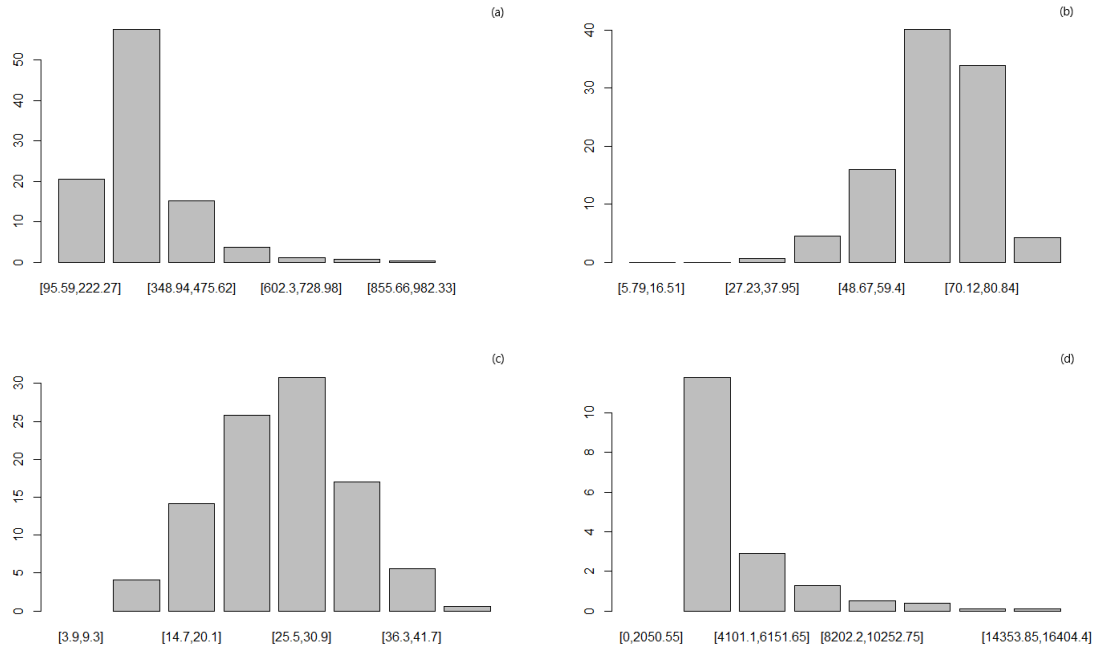
Fonte: Elaborada pela autora.

Tabela 32 – Avaliação do desempenho dos modelos para a região do Nordeste brasileiro.

Modelos	**(Min; Max)**	**(Cen; Amp)**
Regressão linear intervalar	0,2116	0,1244
Regressão car intervalar	0,2102	0,1259

Fonte: Elaborada pela autora.

Figura 24 – Histogramas para variáveis intervalares do Nordeste brasileiro: renda per capita (a); percentual de famílias que ganham meio salário mínimo (b), taxa de analfabetismo (c); densidade demográfica (d).



Fonte: Elaborada pela autora.

4.5 CONSIDERAÇÕES FINAIS

Neste capítulo, apresentamos um novo modelo de regressão para dados espaciais intervalares na estrutura de SDA. Mais precisamente, introduzimos um modelo que agrega informação espacial nas variáveis intervalares. Em particular, o objetivo principal dessa metodologia consiste em levar em conta a variabilidade presente nas variáveis intervalares no contexto da análise espacial.

Realizamos estudos simulados que ilustram o comportamento do modelo com efeitos globais contendo dependência espacial em diferentes cenários. Os resultados mostraram que o modelo de regressão espacial intervalar apresenta um desempenho satisfatório. Aplicações para conjuntos de dados reais foram consideradas. Tratamos modelos de regressão espacial intervalar para o Estado de Pernambuco e para a região Nordeste do Brasil. Observamos que em áreas com menor variabilidade espacial (Aplicação I), o modelo de regressão espacial intervalar tem melhor desempenho que os métodos usuais. Ao considerar áreas com maior variabilidade espacial (Aplicação II), ambos os métodos apresentam resultados semelhantes. Então, os métodos estatísticos espaciais aplicados neste trabalho foram eficientes para modelar dados espaciais em que as variáveis analisadas são do tipo intervalar.

5 CONCLUSÕES

Os índices de autocorrelação espacial são bastantes utilizados na literatura com a finalidade de verificar as dependências espaciais, buscando identificar padrões e níveis de agrupamento. Dentre os índices de autocorrelação espacial, o Índice de Moran é o mais utilizado até o momento na literatura. Até então, apenas estudos utilizando variáveis clássicas tinham sido abordados. Dessa forma, é importante pensarmos em generalizações para tais índices.

Diante desse fato, neste estudo apresentamos um novo índice de autocorrelação para dados de intervalo no framework de SDA. Este índice é uma extensão do índice de Moran para dados de intervalo. Nós vimos que a abordagem proposta nos permitiu assumir duas formas para o cálculo da matriz de pesos. A primeira forma consiste em adotar as coordenadas de latitude e longitude descritas por valores centrais, como é usado com dados clássicos. A segunda apresenta uma nova abordagem, em que assumimos variáveis intervalares para descrever as coordenadas de latitude e longitude. Além disso, usamos distâncias para dados de intervalo para identificar a região de vizinhança. A ideia desta segunda abordagem é levar em consideração a variabilidade presente na variável intervalar, bem como a variabilidade presente nas coordenadas geográficas. Propusemos esta abordagem devido ao fato da própria variável intervalar incluir informações sobre a variabilidade. Desse modo, conseguimos determinar a variabilidade das coordenadas geográficas em uma dada área no espaço.

Os estudos de simulação ilustraram o comportamento das abordagens propostas sob diferentes cenários e com diferentes parametrizações. Os resultados mostraram que o índice de autocorrelação espacial introduzido foi capaz de capturar a autocorrelação existente nos dados. Na maioria dos casos, como esperado, o valor do índice intervalar é inferior ao índice clássico. Aplicações para conjuntos de dados reais intervalares foram consideradas e os resultados mostram a importância do índice proposto em relação aos diferentes critérios para o cálculo da matriz de pesos, com base em distâncias intervalares. Isso se justifica pelo fato de que quando trabalhamos com dados de intervalos, não temos o valor exato da observação, entretanto, o índice foi capaz de obter resultados semelhantes. Na prática, isso pode ser útil para situações em que não temos acesso a um conjunto de dados detalhado de varáveis clássicas, mas sim, intervalares. Com base na análise exploratória, buscamos identificar os comportamentos das informações georreferenciadas para dados simbólicos do tipo intervalar.

Modelos de regressão no contexto de dados simbólicos do tipo intervalo foram utilizados.

O Modelo CAR que incorpora a dependência espacial também foi considerado. Observamos o desempenho desses modelos quando os dados são autocorrelacionados tanto no modelo de regressão linear para dados do tipo intervalo, quanto o modelo CAR no contexto intervalar. De acordo com essses resultados, verificamos que esses modelos tendem a ter um desempenho superior quando existe uma forte autocorrelação espacial nos dados. Além disso, quando utilizamos o CAR intervalar com distâncias intervalares, observamos uma homogeneidade nos valores para o desempenho dos modelos. De maneira geral, não encontramos diferenças entre as parametrizações utilizadas para os modelos analisados.

Quando comparamos o modelo de regressão linear para dados do tipo intervalo com o modelo de regressão para dados espaciais com a matriz de vizinhança calculada através de distâncias para pontos, observamos semelhanças no desempenho. A identificação do modelo que apresenta o melhor desempenho está relacionada a diversos fatores, como a variabilidade existente, o valor de (ρ), o cálculo da matriz de vizinhança, entre outros.

De maneira geral, os métodos estatísticos espaciais aplicados para dados simbólicos do tipo intervalo apresentaram-se satisfatórios na quantificação da autocorrelação espacial e na aplicação dos modelos analisados.

Os conhecimentos obtidos através do desenvolvimento desta tese podem ser consideravelmente ampliados para lidar com problemas no qual temos a informação georreferenciada no contexto de SDA. Para a continuidade do trabalho de pesquisa descrito neste documento, listamos algumas propostas de trabalhos futuros a serem realizados.

1. Ampliar a literatura de Análise Espacial com a implementação de outros métodos de identificação de autocorrelação espacial no contexto de dados simbólicos;

2. Aplicar outras métricas para a avaliação de desempenho dos modelos em que possa ser utilizada uma ponderação em relação ao tamanho das áreas;

3. Estender o nosso estudo para o caso do Indicador de Associação Espacial Local- LISA.

REFERÊNCIAS

ANSELIN, L. Local indicators of spatial association—lisa. *Geographical analysis*, Wiley Online Library, v. 27, n. 2, p. 93–115, 1995.

BAILEY, T. C.; GATRELL, A. C. *Interactive spatial data analysis*. [S.l.]: Longman Scientific & Technical Essex, 1995. v. 413.

BANERJEE, S.; CARLIN, B. P.; GELFAND, A. E. *Hierarchical modeling and analysis for spatial data*. [S.l.]: CRC press, 2014.

BAO, C.; PENG, H.; HE, D.; WANG, J. Adaptive fuzzy c-means clustering algorithm for interval data type based on interval-dividing technique. *Pattern Analysis and Applications*, Springer, v. 21, n. 3, p. 803–812, 2018.

BASU, S.; THIBODEAU, T. G. Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, Springer, v. 17, n. 1, p. 61–85, 1998.

BERTRAND, P.; GOUPIL, F. Descriptive statistics for symbolic data. In: *Analysis of symbolic data*. [S.l.]: Springer, 2000. p. 106–124.

BESAG, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 36, n. 2, p. 192–225, 1974.

BILLARD, L.; DIDAY, E. Regression analysis for interval-valued data. In: *Data Analysis, Classification, and Related Methods*. [S.l.]: Springer, 2000. p. 369–374.

BILLARD, L.; DIDAY, E. Symbolic regression analysis. In: *Classification, Clustering, and Data Analysis*. [S.l.]: Springer, 2002. p. 281–288.

BILLARD, L.; DIDAY, E. Descriptive statistics for interval-valued observations in the presence of rules. *Computational Statistics*, Springer, v. 21, n. 2, p. 187–210, 2006.

BILLARD, L.; DIDAY, E. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. [S.l.]: John Wiley, 2006.

BILLARD, L.; DIDAY, E. *Symbolic Data Analysis: Conceptual Statistics and Data Mining John Wiley*. [S.l.]: Chichester, 2006.

BIVAND, R. S.; PEBESMA, E. J.; GOMEZ-RUBIO, V.; PEBESMA, E. J. *Applied spatial data analysis with R*. [S.l.]: Springer, 2008. v. 747248717.

BIVAND, R. S.; WONG, D. W. S. Comparing implementations of global and local indicators of spatial association. *TEST*, Taylor and Francis, v. 27, p. 716–748, 2018.

BOCK, H.-H.; DIDAY, E. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. [S.l.]: Springer Science & Business Media, 1999.

BRAZ, G. J.; CARDOSO, A. de P.; CORRÊA, A. S.; CESAR, A. M. de O. Classification of breast tissues using moran's index and geary's coefficient as texture signatures and svm. *Computers in biology and medicine*, v. 39, n. 12, p. 1063–1072, 2009.

CÂMARA, G.; MONTEIRO, A. M.; FUCKS, S. D.; CARVALHO, M. S. Análise espacial e geoprocessamento. *Análise espacial de dados geográficos. Brasília: EMBRAPA*, p. 21–54, 2004.

CARVALHO, F. d. A. D.; NETO, E. d. A. L.; FERREIRA, M. R. A robust regression method based on exponential-type kernel functions. *Neurocomputing*, Elsevier, v. 234, p. 58–74, 2017.

CARVALHO, F. d. A. D.; SOUZA, R. M. de; CHAVENT, M.; LECHEVALLIER, Y. Adaptive hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, Elsevier, v. 27, n. 3, p. 167–179, 2006.

CARVALHO, F. d. A. de. Histograms in symbolic data analysis. *Annals of Operations Research*, Springer, v. 55, n. 2, p. 299–322, 1995.

CARVALHO, F. d. A. de; SIMÕES, E. C. Fuzzy clustering of interval-valued data with city-block and hausdorff distances. *Neurocomputing*, Elsevier, v. 266, p. 659–673, 2017.

CHAVENT, M.; LECHEVALLIER, Y. Dynamical clustering of interval data: Optimization of an adequacy criterion based on hausdorff distance. In: *Classification, clustering, and data analysis*. [S.l.]: Springer, 2002. p. 53–60.

CHEN, Y. New approaches for calculating moran's index of spatial autocorrelation. *PloS one*, Public Library of Science, v. 8, n. 7, p. e68336, 2013.

CIMA, E. G.; URIBE-OPAZO, M. A.; JOHANN, J. A.; JR, W. F. d. R.; DALPOSSO, G. H. Analysis of spatial autocorrelation of grain production and agricultural storage in parana. *Engenharia Agrícola*, SciELO Brasil, v. 38, n. 3, p. 395–402, 2018.

CRESSIE, N. Statistics for spatial data. *Terra Nova*, Wiley Online Library, v. 4, n. 5, p. 613–617, 1992.

CRESSIE, N.; CHAN, N. H. Spatial modeling of regional variables. *Journal of the American Statistical Association*, Taylor & Francis, v. 84, n. 406, p. 393–401, 1989.

DIDAY, E. Thinking by classes in data science: the symbolic data analysis paradigm. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 8, n. 5, p. 172–205, 2016.

DIDAY, E. Improving explanatory power of machine learning in the symbolic data analysis framework. In: SPRINGER. *International Workshop on Artificial Intelligence and Pattern Recognition*. [S.l.], 2018. p. 3–14.

DIDAY, E.; NOIRHOMME-FRAITURE, M. *Symbolic data analysis and the SODAS software*. [S.l.]: John Wiley & Sons, 2008.

DOMINGUES, M. A.; SOUZA, R. M. de; CYSNEIROS, F. J. A. A robust method for linear regression of symbolic interval data. *Pattern Recognition Letters*, Elsevier, v. 31, n. 13, p. 1991–1996, 2010.

DORMANN, C. F.; MCPHERSON, J. M.; ARAÚJO, M. B.; BIVAND, R.; BOLLIGER, J.; CARL, G.; DAVIES, R. G.; HIRZEL, A.; JETZ, W.; KISSLING, W. D. et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, Wiley Online Library, v. 30, n. 5, p. 609–628, 2007.

DOUZAL-CHOUAKRIA, A.; BILLARD, L.; DIDAY, E. Principal component analysis for interval-valued observations. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Wiley Online Library, v. 4, n. 2, p. 229–246, 2011.

EFRON, B. Bootstrap methods: another look at the jackknife. e annals of statistics, 7 (1): 1–26. *URL http://www. jstor. org/stable/2958830*, 1979.

EFRON, B. Bootstrap methods: another look at the jackknife. In: *Breakthroughs in statistics*. [S.l.]: Springer, 1992. p. 569–593.

ER, A. C.; ROSLI, M.; ASMAHANI, A.; NAIM, M. M.; HARSUZILAWATI, M. Spatial mapping of dengue incidence: A case study in hulu langat district, selangor, malaysia. *International Journal of Human and Social Sciences*, v. 5, n. 6, p. 3410–3414, 2010.

FAGUNDES, R. A.; SOUZA, R. M. D.; CYSNEIROS, F. J. A. Robust regression with application to symbolic interval data. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 26, n. 1, p. 564–573, 2013.

FAGUNDES, R. A.; SOUZA, R. M. D.; CYSNEIROS, F. J. A. Interval kernel regression. *Neurocomputing*, Elsevier, v. 128, p. 371–388, 2014.

FAGUNDES, R. A.; SOUZA, R. M. de; SOARES, Y. M. Quantile regression of interval-valued data. In: IEEE. *2016 23rd International Conference on Pattern Recognition (ICPR)*. [S.l.], 2016. p. 2586–2591.

FISCHER, M. M.; WANG, J. *Spatial data analysis: models, methods and techniques*. [S.l.]: Springer Science & Business Media, 2011.

FREITAS, W. W.; SOUZA, R. M. de; AMARAL, G. J.; BASTIANI, F. D. Exploratory spatial analysis for interval data: A new autocorrelation index with covid-19 and rent price applications. *Expert Systems with Applications*, Elsevier, v. 195, p. 116561, 2022.

GETIS, A.; BOOTS, B. N. *Models of spatial processes: an approach to the study of point, line and area patterns*. [S.l.]: Cambridge University Press, 1978. v. 198.

GETIS, A.; ORD, J. 1992: The analysis of spatial association by use of distance statistics. geographical analysis 24, 189-206. 1992.

GOODCHILD, M. F. *Spatial autocorrelation*. [S.l.]: Geo Books, 1986. v. 47.

HAINING, R. P.; HAINING, R. *Spatial data analysis: theory and practice*. [S.l.]: Cambridge University Press, 2003.

HAO, P.; GUO, J. Constrained center and range joint model for interval-valued symbolic data regression. *Computational Statistics & Data Analysis*, Elsevier, v. 116, p. 106–138, 2017.

HOEF, J. M. V.; HANKS, E. M.; HOOTEN, M. B. On the relationship between conditional (car) and simultaneous (sar) autoregressive models. *Spatial statistics*, Elsevier, v. 25, p. 68–85, 2018.

KAO, C.-H.; NAKANO, J.; SHIEH, S.-H.; TIEN, Y.-J.; WU, H.-M.; YANG, C.-K.; CHEN, C.-h. Exploratory data analysis of interval-valued symbolic data with matrix visualization. *Computational Statistics & Data Analysis*, Elsevier, v. 79, p. 14–29, 2014.

LAURO, N.; VERDE, R.; PALUMBO, F. Factorial methods with cohesion constraints on symbolic objects. In: *Data Analysis, Classification, and Related Methods*. [S.l.]: Springer, 2000. p. 381–386.

LI, W.; GUO, J.; CHEN, Y.; WANG, M. A new representation of interval symbolic data and its application in dynamic clustering. *Journal of Classification*, Springer, v. 33, n. 1, p. 149–165, 2016.

LIM, C. Interval-valued data regression using nonparametric additive models. *Journal of the Korean Statistical Society*, Springer, v. 45, p. 358–370, 2016.

LIMA, L. M. M. de; SÁ, L. R. de; MACAMBIRA, A. F. U. dos S.; NOGUEIRA, J. de A.; VIANNA, R. P. de T.; MORAES, R. M. de. A new combination rule for spatial decision support systems for epidemiology. *International journal of health geographics*, Springer, v. 18, n. 1, p. 1–10, 2019.

MATHERON, G. Principles of geostatistics. *Economic geology*, Society of Economic Geologists, v. 58, n. 8, p. 1246–1266, 1963.

MONTEIRO, A. M. V.; CÂMARA, G.; CARVALHO, M.; DRUCK, S. Análise espacial de dados geográficos. *Brasília: Embrapa*, 2004.

MORAN, P. A. Notes on continuous stochastic phenomena. *Biometrika*, JSTOR, v. 37, n. 1/2, p. 17–23, 1950.

NEGREIROS, J. G.; PAINHO, M.; AGUILAR, F. J.; AGUILAR, M. A. A comprehensive framework for exploratory spatial data analysis: Moran location and variance scatterplots. *International Journal of Digital Earth*, Taylor & Francis, v. 3, n. 2, p. 157–186, 2010.

NETO, E. d. A. L.; CARVALHO, F. d. A. de. Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis*, Elsevier, v. 52, n. 3, p. 1500–1515, 2008.

NETO, E. d. A. L.; CARVALHO, F. d. A. de. Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis*, Elsevier, v. 54, n. 2, p. 333–347, 2010.

NETO, E. d. A. L.; CARVALHO, F. d. A. de. Nonlinear regression applied to interval-valued data. *Pattern Analysis and Applications*, Springer, v. 20, n. 3, p. 809–824, 2017.

NETO, E. d. A. L.; CORDEIRO, G. M.; CARVALHO, F. d. A. de. Bivariate symbolic regression models for interval-valued variables. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 81, n. 11, p. 1727–1744, 2011.

NOIRHOMME-FRAITURE, M. Visualization of large data sets: the zoom star solution. *International Electronic Journal of Symbolic Data Analysis*, Citeseer, p. 26–39, 2002.

OBAMA, M. *Minha história*. [S.l.]: Objetiva, 2018.

OCHS, M.; DIDAY, E.; AFONSO, F. From the symbolic analysis of virtual faces to a smiles machine. *IEEE transactions on cybernetics*, IEEE, v. 46, n. 2, p. 401–409, 2015.

OLIVEIRA, M. R.; AZEITONA, M.; PACHECO, A.; VALADAS, R. Population symbolic covariance matrices for interval data. *arXiv preprint arXiv:1810.06474*, 2018.

ORD, J. K.; GETIS, A. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis*, Wiley Online Library, v. 27, n. 4, p. 286–306, 1995.

PEDROSA, N. L.; ALBUQUERQUE, N. L. S. d. Spatial analysis of covid-19 cases and intensive care beds in the state of ceará, brazil. *Ciência & Saúde Coletiva*, SciELO Public Health, v. 25, p. 2461–2468, 2020.

PIMENTEL, B. A.; SOUZA, R. M. de. A weighted multivariate fuzzy c-means method in interval-valued scientific production data. *Expert Systems with Applications*, Elsevier, v. 41, n. 7, p. 3223–3236, 2014.

REYES, D. M.; SOUZA, R. M. de; OLIVEIRA, A. L. de. A three-stage approach for modeling multiple time series applied to symbolic quartile data. *Expert Systems with Applications*, Elsevier, p. 115884, 2021.

REYES, D. M. A.; SOUZA, R. M. C. R. D.; CYSNEIROS, F. J. D. A. Predicting symbolic interval-valued data through symmetrical nonlinear regression. *International Journal of Business Intelligence and Data Mining*, Inderscience Publishers (IEL), v. 12, n. 2, p. 175–189, 2017.

RIGBY, R. A.; STASINOPOULOS, M. D.; HELLER, G. Z.; BASTIANI, F. D. *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. [S.l.]: CRC press, 2019.

RIPLEY, B. *198 1. Spatial Statistics*. [S.l.]: Wiley, New York, 1984.

RIZZO, M. L. *Statistical computing with R*. [S.l.]: CRC Press, 2019.

ROGERS, A. *Statistical analysis of spatial dispersion: the quadrat method*. [S.l.]: Pion, 1974. v. 6.

SCHABENBERGER, O.; GOTWAY, C. A. *Statistical methods for spatial data analysis*. [S.l.]: CRC press, 2017.

SILVA, W. J.; SOUZA, R. M.; CYSNEIROS, F. J. A. Polygonal data analysis: A new framework in symbolic data analysis. *Knowledge-Based Systems*, Elsevier, v. 163, p. 26–35, 2019.

SNOW, J.; RICHARDSON, B. W.; FROST, W. H. *Snow on cholera*. [S.l.]: Commonwealth Fund New York, 1936.

SOKAL, R. R.; THOMSON, B. A. Population structure inferred by local spatial autocorrelation: an example from an amerindian tribal population. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, Wiley Online Library, v. 129, n. 1, p. 121–131, 2006.

SOUZA, L. C.; SOUZA, R. M.; AMARAL, G. J.; FILHO, T. M. S. A parametrized approach for linear regression of interval data. *Knowledge-Based Systems*, Elsevier, v. 131, p. 149–159, 2017.

SOUZA, R. M. C. de; CARVALHO, F. d. A. de; TENÓRIO, C. P.; LECHEVALLIER, Y. Dynamic cluster methods for interval data based on mahalanobis distances. In: *Classification, Clustering, and Data Mining Applications*. [S.l.]: Springer, 2004. p. 351–360.

SOUZA, R. M. de; CARVALHO, F. d. A. D. Clustering of interval data based on city-block distances. *Pattern Recognition Letters*, Elsevier, v. 25, n. 3, p. 353–365, 2004.

SURESH, N.; BRITO, P.; DIAS, S. Prediction of pollution levels from atmospheric variables a study using clusterwise symbolic regression. 2020.

TANGO, T. A class of tests for detecting 'general'and 'focused'clustering of rare diseases. *Statistics in Medicine*, Wiley Online Library, v. 14, n. 21-22, p. 2323–2334, 1995.

TOBLER, W. R. A computer movie simulating urban growth in the detroit region. *Economic geography*, Taylor & Francis, v. 46, n. sup1, p. 234–240, 1970.

TSAI, P.-J.; PERNG, C.-H. Spatial autocorrelation analysis of 13 leading malignant neoplasms in taiwan: a comparison between the 1995-1998 and 2005-2008 periods. *Health*, Scientific Research Publishing, v. 3, n. 12, p. 712, 2011.

WALL, M. M. A close look at the spatial correlation structure implied by the car and sar models. *Res. Rep*, n. 2000-022, 2000.

WALLER, L. A.; GOTWAY, C. A. *Applied spatial statistics for public health data,*. [S.l.]: John Wiley & Sons, 2004. v. 368.

WORLD HEALTH ORGANIZATION. Coronavirus disease 2019 (covid-19): situation report, 46. *Coronavirus disease 2019 (COVID-19): Situation Report*, World Health Organization, v. 46, 2020. Disponível em: <https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf?sfvrsn=96b04adf_4>.

XU, W. *Symbolic data analysis: interval-valued data regression*. Tese (Doutorado) — University of Georgia Athens, GA, 2010.

YANG, M.-S.; HUNG, W.-L.; CHEN, D.-H. Self-organizing map for symbolic data. *Fuzzy Sets and Systems*, Elsevier, v. 203, p. 49–73, 2012.

APÊNDICE A – ALGUMAS DEFINIÇÕES

A.1 DEFINIÇÃO

A relação cartesiana $A \oplus B$ entre dois conjuntos A e B é a união de seus componentes,

$$A \oplus B = (A_1 \oplus B_1, \cdots A_p \oplus B_p)$$

onde $A_j \oplus B_j = {}'A_j \cup B'_j$. Quando A e B são objetos de valor de intervalo com $A_j = [a_j^A, b_j^A]$ e $B_j = [a_j^B, b_j^B]$, então

$$A \oplus B = [Min(a_j^A \cdot a_j^B), Max(b_j^A \cdot b_j^B)]. \tag{A.1}$$

A.2 DEFINIÇÃO

O encontro cartesiano $A \otimes B$ entre dois conjuntos A e B é a interseção dos componentes,

$$A \otimes B = (A_1 \otimes B_1, \cdots, A_p \otimes B_p)$$

onde, $A_j \otimes B_j = A_j \cap B_j$, $j = 1, \cdots, p$. Quando A e B são objetos com valores de intervalo formando intervalos sobrepostos em Y_j,

$$A_j \otimes B_j = [Max(a_j^A \cdot a_j^B), Min(b_j^A \cdot b_j^B)], \tag{A.2}$$

e quando $A_j \cap B_j = \phi$, então $A_j \otimes B_j = 0$.

APÊNDICE B – PUBLICAÇÕES

Este apênndice apresenta as publicações decorrentes da pesquisa desenvolvida ao longo do curso de doutorado, para a confecção desta tese. Dois trabalhos foram publicados em congressos da área de Estatística, sendo um pôster e uma apresentação oral. Um artigo foi publicado em periódico internacional. Um outro artigo encontra-se submetido.

- FREITAS, W. W. L.; Reyes, D.M.A.; SOUZA, R. M. C. R.; Cysneiros, F. J. A.. *An interval Approach to Risk Estimation in Determining Capital Assets.* In: 18ª ESTE - Escola de Séries Temporais e Econometria, 2019, Gramado-RS.

- FREITAS, WANESSA W.L.; DE SOUZA, RENATA M.C.R.; AMARAL, GETÚLIO J.A.; DE BASTIANI, FERNANDA. *A Spatial Autocorrelation Index for Symbolic Interval Data.* SINAPE, 2022, Gramado-RS.

- FREITAS, WANESSA W.L.; DE SOUZA, RENATA M.C.R.; AMARAL, GETÚLIO J.A.; DE BASTIANI, FERNANDA. *Exploratory spatial analysis for interval data: A new autocorrelation index with COVID-19 and rent price applications.* EXPERT SYSTEMS WITH APPLICATIONS, v. n/a, p. 116561, 2022.

Figura 25 – Publicação resultante dessa tese.

Expert Systems With Applications 195 (2022) 116561

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Exploratory spatial analysis for interval data: A new autocorrelation index with COVID-19 and rent price applications

Wanessa W.L. Freitas [a], Renata M.C.R. de Souza [a,*], Getúlio J.A. Amaral [b], Fernanda De Bastiani [b]

[a] Universidade Federal de Pernambuco, Centro de Informática, Av. Jornalista Aníbal Fernandes, s/n – Cidade Universitária, 50.740-560 Recife, PE, Brazil
[b] Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Departamento de Estatística, Av. Jornalista Aníbal Fernandes, s/n - Cidade Universitária, 50740-560, Recife, PE, Brazil

ARTICLE INFO

Keywords:
Interval data
Moran's index
Spatial analysis
Symbolic data analysis
Spatial variability

ABSTRACT

This paper aims to identify the behavior of interval data associated to its respective geospatial information with in the framework of Symbolic Data Analysis. The main idea is to extend Moran's autocorrelation index of Exploratory Spatial Analysis to interval data. Symbolic data analysis is a domain of research and application related to the areas of machine learning and statistics that provide tools to describe units (objects), enabling them to consider variability. Spatially correlated data are geospatial data with spatial autocorrelation, and the variability that comes from each region and neighborhood may be better expressed by intervals. Thus, this paper demonstrates the importance of considering the variability present in the interval variable and the variability present in geographical information. Experiments with synthetic interval data are performed to illustrate the usefulness of the proposed approach. We also, analyze two applications, dealing with COVID-19 and rent price interval data.

1. Introduction

The dynamics of the phenomena that occur in geographical space constitute one of the challenges for clarifying central issues in several areas of knowledge. For instance, point pattern analysis (see Getis and Boots (1978), Ripley (1984) and Rogers (1974)) is a set of techniques with enormous relevance in describing geographical phenomena. Spatial Data Analysis is concerned with questions that may not be answered only by direct observations of a data set. In this type of analysis, it is also necessary to consider the behavior of a data set by taking into account its spatial characteristics. Spatial data analysis provides methods for creating and interpreting maps, and is crucial for assisting in decision-making (Bivand, Pebesma, Gomez-Rubio, & Pebesma, 2008).

One of the essential ideas in spatial data analysis is the concept of spatial autocorrelation. This quantity measures the similarity between objects or activities in a given geographic area, based on their proximity (Fischer & Wang, 2011). It is possible to note its existence from the proposition of Tobler (1970) called 'the first law of geography: everything is related to everything else, but close things are more related than distant things'. In this context, for a given observed event, there is a corresponding spatial location, which may be described by its position on a map or by geographic coordinates in a reference system (Bivand et al., 2008). It is possible to detect the presence of spatial autocorrelation by using several techniques which deal simultaneously with location and attribute information. In general, spatial autocorrelation compares two sets of similarities (Cressie, 1992).

More specifically, the main goal of calculating spatial autocorrelation is to verify the existence of a spatial pattern. It is possible to describe the behavior of a random variable by considering the geographic space where the phenomena of interest occur. Measures of spatial autocorrelation may be helpful for indicating the degree of similarity between different regions. One example, of this would be, a situation in which we study indexes related to agricultural production in neighboring areas that historically have presented similar behavior. We then, repeat the analysis in a given period and observe an abrupt change in this behavior, now indicating the absence of autocorrelation. This could occur because of production problems in one of the regions, such as weather, or pests on farms. Hence, spatial autocorrelation may therefore assist in making decisions to solve these problems and provide a quantifier to compare the production performance between the regions analyzed.

Studies involving the concept of spatial autocorrelation have been reported in several areas. Recently, a study on soybean production in the Brazilian State of Paraná, identified regions with low and high

* Corresponding author.
E-mail addresses: wwlf@cin.ufpe.br (W.W.L. Freitas), rmcrs@cin.ufpe.br (R.M.C.R. de Souza), gjaa@de.ufpe.br (G.J.A. Amaral), debastiani@de.ufpe.br (F. De Bastiani).

https://doi.org/10.1016/j.eswa.2022.116561
Received 6 September 2020; Received in revised form 6 January 2022; Accepted 16 January 2022
Available online 5 February 2022
0957-4174/© 2022 Elsevier Ltd. All rights reserved.

Fonte: Elaborada pela autora.