UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

MultiMapas: abordagem multiobjetivo para construção de mapas coropléticos de dados heterogêneos multifontes

Gesiel Rios Lopes

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Gesiel Rios Lopes

MultiMapas: abordagem multiobjetivo para construção de mapas coropléticos de dados heterogêneos multifontes

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Alexandre Cláudio Botazzo Delbem

USP – São Carlos
Fevereiro de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Lopes, Gesiel Rios
L864m MultiMapas: abordagem multiobjetivo para construção de mapas coropléticos de dados heterogêneos multifontes / Gesiel Rios Lopes; orientador Alexandre Cláudio Botazzo Delbem. -- São Carlos, 2023.
160 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Análise de decisão multicritério. 2. Fusão de Dados. 3. Visualização de Mapas. I. Delbem, Alexandre Cláudio Botazzo , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Gesiel Rios Lopes

MultiMaps: multiobjective approach for building choropleth maps from multisource heterogeneous data

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Alexandre Cláudio Botazzo Delbem

USP – São Carlos
February 2024

Este trabalho é dedicado à minha esposa Karina Jorge Pelarigo e a meu filho Luan Pelarigo Lopes, pela dedicação e compreensão nessa caminhada.

AGRADECIMENTOS

A Deus, criador e doador de toda vida, por iluminar meu caminho, por me guiar e por não me deixar cair perante todas as dificuldades enfrentadas.

A minha querida esposa Karina, por ter me mostrado um lado da vida que não conhecia, por todo apoio, carinho, incentivo, companheirismo e amor. Obrigado pelo carinho, compreensão e paciência nos inúmeros momentos difíceis, por sempre ficar ao meu lado em todos os momentos e pelo Luan, nosso filho tão desejado e amado.

Ao meu orientador, Alexandre Delbem por ter me acolhido quando necessitei e pela sua dedicação durante o desenvolvimento deste trabalho. Sua educação e paciência na forma de lapidar esta pesquisa, quando estava em andamento, foram fundamentais para minha formação. Após cada conversa, sempre saía tendo a certeza de ter crescido profissional e pessoalmente.

Ao Alessandro Wilk, um amigo e irmão que a vida me deu, parceiro nas horas boas e difíceis. Obrigado meu irmão por sempre se fazer presente nessa caminhada.

Ao Wellington da Silva Martins, uma pessoa incrível e um grande amigo que fiz nessa jornada.

Ao Roberto Fray pelos ensinamentos, pela disposição e incentivos para seguir em frente.

Agradeço a todos os amigos e amizades construídas no DINTER, mesmo que em poucos momentos juntos, mas foram momentos intensamente verdadeiros.

Aos colegas do LaSDPC, primeiro laboratório que tive o prazer de trabalhar nessa jornada, Leonildo Azevedo, Edvard de Oliveira, Luiz Henrique Nunes, Helder Luz, Lourenço Alves, Hemrique Yoshikazu Shishido, Vinicius Aires, Davi Conti, Guilerme Martins, Leonardo Araruna, Matheus Saldanha, Ana Spengler, Gabriel Tomiatti e demais que conheci.

Aos colegas do LCR Sidgley, Cristiano, Pedro Arantes, Enio Politi, Fernando Elias, Breno Caetano, Celso Lopes, José Luis, Renata Magro, Rubens.

À professora e amiga Mellina Yamamura pela parceria nos artigos publicados e à Denise Scatolini da VIGEP-SC pela disponibilização dos dados e principalmente o seu tempo.

Aos professores do ICMC, em especial, aos Claudio Fabiano Motta Toledo e Solange Oliveira Rezende pelos ensinamentos.

À Secretária de Pós-graduação pela ajuda com os trâmites burocráticos.

Aos órgãos de fomento FAPEMA, FAPESP, CNPq e CAPES pelo apoio financeiro,

possibilitando o desenvolvimento deste trabalho.

E a todos os que, de alguma forma, possibilitaram a realização deste sonho, meu muito obrigado.

"Se fizéssemos todas aquelas coisas de que somos capazes, nós nos surpreenderíamos a nós mesmos."

(Thomas Edison)

RESUMO

LOPES, G. R. **MultiMapas: abordagem multiobjetivo para construção de mapas coropléticos de dados heterogêneos multifontes**. 2024. 160 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

O aumento significativo na disponibilização de dados com informações espaciais, como tempo e espaço, oriundo de diversas fontes têm gerado novas oportunidades de análise e modelagem, buscando um melhor entendimento do sistema associado a esses dados. No entanto, surgem problemas na avaliação das distribuições espaciais e na fusão de dados heterogêneos de múltiplas fontes. Um deles explorado na literatura é identificar os pesos a serem atribuídos a cada entrada de dados. Este trabalho propõe, dessa forma, a utilização de um modelo espacial multicritério que visa otimizar a dependência espacial e a heterogeneidade espacial para pesar o grau de importância de cada componente do modelo e combiná-los, gerando diversos mapas coropléticos de forma automática e sem a necessidade de especialista de domínio de cada fonte de dados. Os resultados, obtidos por meio de estudos de caso, permitiram a comparação da qualidade da solução e do custo computacional das técnicas aplicadas, bem como o desenvolvimento de uma nova técnicas para fusão de dados geoespaciais por meio da análise da dependência e heterogeneidade espacial, além de um arcabouço para o tratamento e manipulação de dados geoespaciais.

Palavras-chave: Análise de decisão multicritério, Fusão de Dados, Visualização de Mapas.

ABSTRACT

LOPES, G. R. **MultiMaps: multiobjective approach for building choropleth maps from multisource heterogeneous data**. 2024. 160 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

The significant increase in the availability of data with spatial information, such as time and space, from different sources generated with new opportunities for analysis and modeling, seeking a better understanding of the system associated with this data. However, problems arise in assessing spatial distributions and merging heterogeneous data from multiple sources. One of the important problems explored in the literature is identifying the weights assigned to each data entry. This work proposes the use of a multi-criteria spatial model that aims to optimize spatial dependence and spatial heterogeneity to weigh the degree of importance of each component of the model and combine them, generating several choropleth maps automatically and without the need for a specialist. domain of each data source. The results obtained through case studies allowed the comparison of the quality of the solution and the computational cost of the applied techniques, as well as the development of a new technique for fusion of geospatial data through the analysis of spatial dependence and heterogeneity, in addition to a framework for the processing and manipulation of geospatial data.

Keywords: Multi-criteria decision analysis, data fusion, map visualization.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADCM	Absolute Deviation Around Class Medians
ADNDP	Ammunution Distribution Network Design Problem
AG	Algoritmos Genéticos
AHP	Analytical Hierarchy Process
ANP	Analytical Network Process
blx-α	*blend alpha crossover*
BWM	Best-Worst Method
COPRAS	Complex Proportional Assessment
GALDIT	Groundwater Aquifer Level Distance Impact Thickness
GIS	Geografic information System
GIScience	Geographic Information Science
IBGE	Instituto Brasileiro de Geografia e Estatística
ICA	Imperial Competitive Algorithm
IDH	Índice de Desenvolvimento Humano
IoT	Internet of Things
IPVS	Índice Paulista de Vulnerabilidade Social
KNN	K-Nearest Neighbors
LiDARs	Light Detection and Ranging
LISA	Local Indicators of Spatial Association
MCDA	multiple-criteria Decision Analysis
MCDM	Multicriteria Deicsion Making
MOEA	Multi-objective Evolutionary Algorithms
MOEA	MultiObjective Evolutionary Algorithms
MOORA	Multi-objective optimization by ratio analysis
MSE	Mapeamento Sistemático de Estudo
MsHD	Multisource Heterogeneous Data
NSGA-II	Non-dominated Sorting Genetic Algorithm II
OGP	Open Government Partnership
PROMETHEE	Preference Ranking Organization METHod for Enrichment of Evaluations
RBF	Radial Basic Function
SAW	Simple Additive Weighting

SEADE	Fundação Sistema Estadual de Análise de Dados
SIG	Sistema de Informação Geográfica
SINAN-Dengue	Sistema de Informação de Agravos de Notificação
SIVEP-gripe	Sistemas de Informação de Vigilância Epidemiológica da Gripe
SMS	Systematic Mapping Study
SWARA	Stepwise Weight Assessment Ratio Analysis
TBWeb	Sistema de Controle de Pacientes com Tuberculose
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution
USGS	United States Geological Survey
VIGEP-SP	Vigilância Epidemiológica de São Carlos
WLC	Weighted Linear Combination
WLC	weighted linear combination
WOA	Weighted Overlay Analysis

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Informações digitais, comuns em aplicações de elétrica e eletrônica, têm crescido também em outros domínios com base nos avanços em tecnologias de aquisição de dados, por exemplo, como no caso de Internet das Coisas (*Internet of Things (IoT)*). Alguns exemplos mais atuais em estágio de construção ou ampliação, são as redes veiculares (usando sensores de detecção por pulsos de luz infravermelho (*Light Detection and Ranging (LiDARs)*), radares, câmeras, e sensores de fricção), o sensoriamento de clima (por meio de satélites, pluviômetros e radares), o monitoramento de florestas, sistemas de informação de saúde, bases de dados sociais (sobre renda, habitação, poluição, segurança, escolaridade, empregabilidade, transporte e lazer), cujos acessos a eles têm sido melhorado por iniciativas como a Parceria para Governo Aberto (*Open Government Partnership (OGP)*).

Em outras palavras, Dados Heterogêneos Multifontes (*Multisource Heterogeneous Data (MsHD)*) estão crescendo para várias áreas de aplicação. Além disso, muitas dessas áreas são multidisciplinares; com as que envolvem dinâmicas populacionais. Por exemplo, enchentes geralmente se relacionam com eventos climáticos extremos e uma ocupação indevida do solo. De certa forma, MsHD podem tanto ter função complementar, ao melhorarem a qualidade das informações (*e.g.*, estimativas para redes veiculares podem ter mais acurácia com o uso de mais sensores), quanto podem ser essenciais para a obtenção de um modelo inicial (*e.g.*, a exposição a um evento extremo deve utilizar informações sociais).

Este trabalho foca em MsHD multidisciplinares e associados a populações. Nesse contexto, uma forma de modelagem conveniente tem sido a construção de mapas coropléticos (um tipo de mapa temático usado para representar um fenômeno geográfico). Tais informações contidas nos mapas podem reduzir a complexidade inerente desses sistemas o que beneficia o entendimento humano. Tal característica é mais relevante em problemas multidisciplinares, uma vez que envolvem equipes de trabalho com formação heterogênea, em que a linguagem dos mapas georreferenciados facilita a comunicação.

A construção de tais mapas, em geral, é um trabalho árduo, de longo prazo, e, muitas vezes, caro, pois requer a coleta e a sistematização de informações das multifontes e o trabalho conjunto de equipes com especialistas das várias áreas envolvidas. Uma maneira de superar esses obstáculos é pela construção de forma automática de mapas coropléticos a partir de MsHD. Tal automação pode ser obtida a partir de técnicas de aprendizado de máquina, a qual gerará mapas de importância social e econômica, principalmente em situações de eventos críticos, em que há relativamente pouco tempo para a tomada de decisão.

Sendo que a importância de MsHD para modelagem de sistemas complexos foi identificada há certo tempo, conforme ilustra o processo de aprendizado de redes neurais (NEDELJKOVIC; MILOSAVLJEVIC, 1992). O desenvolvimento de métodos para integração e fusão de MsHD é mais recente (JIANG; SHENG; YANG, 2011; LU *et al.*, 2012; LI; PAN, 2012; DEVI *et al.*, 2013; DING; WU *et al.*, 2013; CAO; MIAO; YANG, 2013). As investigações escalaram com os sistemas de *Big Data* (HUANG *et al.*, 2014), atingindo milhares de propostas *ad hoc* (para fins estritos) até 2023.

Em geral, esses trabalhos focam em técnicas de fusão de dados em IoT e/ou em métodos para tomada de decisão estendidos para MsHD.

Outro ponto a se destacar é que a análise de MsHD com a construção de mapas coropléticos ou temáticos é um problema complexo (HE; XU; JIANG, 2022) com menos estudos encontrados na literatura. Uma vez que o processo usual de construção de mapas depende de conhecimento de aspectos práticos do sistema envolvido, como a importância relativa de fatores, pontos de referência ou interesse, que balizam a qualidade dos modelos gerados, em outras informações. Em geral, esse conhecimento não está disponível para MsHD com fontes de diferentes áreas.

Este trabalho propõe, desta forma, um método de inteligência artificial que busca vencer essa barreira. O método é baseado em aprendizado não supervisionado, uma vez que em problemas complexos relativamente novos é raro a disponibilidade de dados com rotulação adequada em termos da representatividade das configurações possíveis para o sistema. Nem por isso, o método deixa de ter uma referência para medição da qualidade dos modelos (mapas) construídos, visto que contorna a falta de rotulações pelo uso das métricas geoespaciais de consistência espaço-temporal disponíveis na literatura para orientar um algoritmo de otimização multiobjetivo.

Destaca-se que a utilização de mais de um critério para guiar o processo de construção do mapa é importante não somente para que o aprendizado possa ser não supervisionado. Os múltiplos critérios/métricas de avaliação do mapa evitam: *i*) efeitos "colaterais" de um possível viés da métrica no modelo gerado, *ii*) a escolha de uma das métricas em detrimento das outras ou de pesos para ponderação entre elas, e *iii*) a determinação de pesos para combinar as várias camadas de informação disponibilizadas pelas multifontes. É de relevância também notar que definições de pesos, em geral, demandam um especialista ou um consenso entre especialistas de

várias áreas, o que pode ser inviável em problemas multidisciplinares.

Como ponto positivo também devemos evidenciar que o método, denominado MultiMapas, foi testado em MsHD da área de Vigilância em Saúde (envolvendo múltiplos agravos) e na área de seguros agricolas. Aplicações em outras áreas são indicadas como exemplo, mas não exploradas em profundidade, buscando ilustrar a flexibilidade do MultiMapas em contribuir em diversas áreas. Esses casos envolvem um ou mais dos seguintes aspectos: aquisição de MsHD requer tempos maiores devido à necessidade de autorização de acesso (como no caso de dados do governo), aprovação de Comitês de Ética e também a digitalização manual para uma ou outra fonte dos dados.

A aplicação do MultiMapas em Vigilância em Saúde, em especial à Dengue (para a qual foi possível acessar MsHD por um período maior), mostra que é possível a construção de mapas coropléticos a partir de MsHD multidisciplinares que são representações visuais consistentes e úteis para tomada de decisão colaborativa.

De certa forma, o MultiMapas é uma proposta que explora tecnologias de Sistema de Informação Geográfica (SIG) (*Geografic information System (GIS)*) para MsHD buscando beneficiar processos de Tomada de Decisão Multicritério (*Multicriteria Deicsion Making (MCDM)*) colaborativa, por equipes multidisciplinares. Assim, o MultiMapas inova ao explorar o conhecimento de três grandes áreas (GIS, MsHD e MCDM) para conseguir de forma automática mapas de síntese confiáveis a partir de dados complexos.

As técnicas de GIS oferecem mecanismos para armazenagem, gestão, análise e visualização de dados geoespaciais para MCDM, o que favorece a percepção de relações espaciais. Isso pode gerar "*insights*", alternativas de decisão e, assim, novas soluções para um problema. Os sistemas envolvendo MsHD trazem novos desafios para o uso de GIS em MCDM que foram enfatizados por demandas de decisões por equipes multidisciplinares. Por exemplo, alguns trabalhos de GIS e MCDM mostraram a necessidade de uma estruturação de informações com relação à granularidade espaço-temporal (MONTEIRO *et al.*, 2004; MALCZEWSKI; RINNER, 2015; DOMINGUES *et al.*, 2020) para geração de modelos mais representativos.

A superação desses desafios pelo MultiMapas envolve a escolha e a integração de métodos de GIS, MsHD e MCDM de forma bem orquestrada. Nesse contexto, a Análise Hierárquica de Processos (*Analytical Hierarchy Process (AHP)*) é um dos métodos fundamentais para tal integração. Outro fator importante é a investigação de Algoritmos Evolutivos MultiObjetivos (*MultiObjective Evolutionary Algorithms (MOEA)*) combinada com as métricas de consistência geoespacial Moran Global (MORAN, 1948), LISA (ANSELIN, 1995) e Geary (GEARY, 1954).

Os mapas gerados são modelos visuais que, apesar de envolver conhecimento de vários áreas, facilitam o entendimento por uma equipe. A construção dos modelos por um MOEA, em geral, produz vários modelos consistentes com as métricas geoespaciais, isso possibilita que a equipe tenha um portfólio de escolhas. Outro aspecto relevante é que a importância relativa de

cada fonte de dados, em cada um dos modelos, pode ser verificada de forma transparente. Em outras palavras, os modelos do MultiMapas são explicáveis, assim, a equipe pode fazer escolhas de maneira mais confiante, ou mesmo, explorar alterações manuais de um modelo.

Por fim, a estratégia de integração do MultiMapas, baseada em MOEAS e AHP, mostra-se também importante para lidar com a complexidade computacional do processo de construção dos mapas coropléticos para MsHD de larga-escala. Para isso, a configuração dos operadores do MOEA é escolhida de acordo com alguns dos modelos teóricos de Computação Evolutiva (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012) que fornecem estimativas de tempo de convergência e de aproximação de soluções ótimas.

Divide-se, portanto, o restante deste texto em mais cinco capítulos, conforme segue:

- No Capítulo 2 é apresentado todo referencial teórico utilizado no escopo dessa pesquisa;
- O Capítulo 3 retrata uma revisão sistemática para levantamento dos principais artigos que utilizam técnicas MCDM, MCDM e algoritmos evolutivos em SIGs;
- No Capítulo 4 são descritas a metodologia proposta nesta tese para gerenciar e integrar dados com características espaciais a partir da otimização da dependência e heterogeneidade espacial;
- O Capítulo 5 descreve os experimentos realizados para validar as técnicas propostas e os resultados obtidos;
- Por fim, o Capítulo 6 contém as conclusões do trabalho, as contribuições, as publicações derivadas da pesquisa, as dificuldades enfrentadas e os trabalhos futuros.

CAPÍTULO

2

REFERENCIAL TEÓRICO

2.1 Considerações iniciais

A compreensão da distribuição espacial, a partir de dados de fenômenos ocorridos no espaço, é um desafio em diversas áreas do conhecimento como na saúde, geologia, agronomia e computação. Contudo, é fundamental entender a distribuição dos dados geoespaciais considerando a localização espacial do fenômeno de forma explícita e traduzi-los em padrões com propriedades mensuráveis e relacionadas. Neste capítulo, apresentam-se os conceitos teóricos essenciais para avançar na compreensão da proposta da tese, assim como de todo o trabalho.

2.2 Dados Heterogêneos de Múltiplas Fontes

Como primeiro conceito, abordar-se-á os dados heterogêneos de múltiplas fontes que são um tipo de dados complexo e diversificado que se origina de múltiplas fontes e apresenta variação significativa em termos de tipos de dados, formatos, qualidade e outras características. Gerir e extrair valor eficazmente desses dados requer uma abordagem multidisciplinar, técnicas avançadas de integração de dados e métodos analíticos robustos (ZHANG, 2010; ZHANG *et al.*, 2018).

Sendo que diferentes fontes de dados fornecem diferentes aspectos do objeto de destino. De modo que, dados heterogêneos de múltiplas fontes podem compensar as deficiências de dados incompletos de uma única fonte de dados, o que torna as informações de destino mais adequadas. Assim, ao eliminar a lacuna entre dados heterogêneos e a fusão de várias fontes de dados para análise de correlação dos dados poderiam emergir novas informações mais valiosas (ZHANG; XING, 2009; BEYER *et al.*, 2010).

Essa fusão de dados é um pré-requisito para a garantia de qualidade e a mineração analítica de dados integrados. No entanto, a fusão de dados na totalidade é um processo de caixa

preta para o usuário, o que torna o processo de fusão de dados carente de interpretabilidade e depuração (ZHANG, 2010; ZHANG *et al.*, 2018).

Segundo Zhang *et al.* (2018), a fusão de heterogêneos multifontes é geralmente dividida em três níveis: 1) fusão da camada de dados; 2) fusão da camada de recursos e 3) fusão da camada de decisão. A fusão da camada de dados é a integração direta dos dados nos dados e análises originais, sendo uma integração de baixo nível. A fusão de camada de recursos é uma fusão de nível médio. Primeiramente, é realizada a extração de características dos dados originais para posteriormente os dados do recurso serem analisados e processados sinteticamente. Esses alcançam uma compressão considerável de informações e facilitam o processamento em tempo real. Já na fusão de camada de tomada de decisão, os dados são processados separadamente de diferentes fontes, o que inclui pré-tratamento, extração de características, identificação ou discriminação, respectivamente, e as conclusões preliminares são obtidas. Em seguida, a decisão de fusão da tomada de decisão é feita através do processamento de correlação e, finalmente, o resultado da inferência conjunta é obtido.

Tal fusão tradicional de informações de múltiplas fontes é um método de processamento de informações para sensores ou sistemas de múltiplas fontes. Ela processa as informações de medição obtidas de múltiplas fontes e usa métodos como associação de informações, integração de informações e filtragem para melhorar a precisão da estimativa do estado alvo e outros recursos e, em última análise, para corrigir a situação, as ameaças e sua avaliação de importância (XIAO-BIN; CHENG-LIN *et al.*, 2005; ZHANG *et al.*, 2018).

Além disso, a fusão de dados de múltiplas fontes constitui de uma ferramenta fundamental em diversas áreas, ao combinar dados heterogêneos de diferentes fontes, como sensores remotos, sistemas de informações geográficas e outras fontes de dados espaciais, os analistas podem obter uma visão mais abrangente de eventos geográficos como mudanças ambientais, movimentos populacionais ou padrões climáticos. A fusão de dados também pode ser usada para melhorar a precisão das estimativas de estados alvo em sistemas de monitoramento espacial (LI, 2020; HE; LIU; DU, 2023).

2.3 Análise Espacial

A partir dos dados espaciais, é possível descobrir não apenas a localização, mais também o comprimento, tamanho, área ou forma de qualquer objeto. Os dados geoespaciais têm inúmeras aplicações em nossa vida cotidiana, indo desde o entendimento e modelagem do comportamento urbano de pessoas, veículos e outros objetos móveis (ZHENG *et al.*, 2014; DOMINGUES *et al.*, 2020) até a Análise Espacial que visa mensurar propriedades e relacionamentos, ao levar em conta a localização espacial do fenômeno em estudo de forma explícita (MONTEIRO *et al.*, 2004).

Segundo Monteiro *et al.* (2004), a taxonomia mais utilizada para caracterizar os proble-

mas de análise espacial considera três tipos de dados:

- *Eventos ou Padrões Pontuais:* fenômenos expressos por meio de ocorrências identificadas como pontos localizados no espaço e no tempo;
- *Superfícies Contínuas:* estimadas a partir de um conjunto de amostras de campo que podem estar regularmente ou irregularmente distribuídas. Normalmente, esse tipo de dado é resultante de levantamento de recursos naturais que incluem mapas geológicos, topográficos, ecológicos, fitogeográficos e pedológicos;
- *Áreas com Contagens e Taxa:* tratam-se de dados associados a levantamentos populacionais, como censos e estatísticas de saúde e que originalmente se referem a indivíduos localizados em pontos específicos do espaço. Esses tipos de dados são normalmente agregados em unidades de análise, usualmente, delimitadas por polígonos fechados como setores censitários, zonas de endereçamento postal, municípios, dentre outros.

A partir desta divisão, verificam-se que os problemas de análise espacial lidam com dados *ambientais* e com dados *socioeconômicos* e em ambos os casos, a análise espacial visa definir um *modelo inferencial* utilizando um conjunto de procedimentos encadeados (MONTEIRO *et al.*, 2004). O processo de modelagem utilizado na análise espacial, em geral, ocorre depois de duas etapas: (i) uma fase de análise exploratória, associada à apresentação visual dos dados sob a forma de gráficos e mapas e (ii) a identificação de padrões de dependência espacial do fenômeno objeto de estudo, sendo este um conceito chave na compreensão e análise dos fenômenos espaciais.

A visualização é provavelmente o método de análise espacial mais comumente usado que resulta em mapas que descrevem padrões espaciais e que são úteis tanto para estimular análises mais complexas quanto para comunicar os resultados de tais análises. Vale ressaltar que um mapa é uma representação do processo geográfico subjacente, mas não é o processo. Apesar do fato de que essas representações não são exatamente corretas em algum sentido, elas são úteis para entender o que é importante sobre um processo geográfico (ALMEIDA, 2012; ANDRADE *et al.*, 2007; CÂMARA *et al.*, 2004).

Já a exploração de dados espaciais envolve o uso de métodos estatísticos para determinar se os padrões observados são aleatórios no espaço. A análise da modelagem introduz o conceito de relações de causa e efeito usando fontes de dados espaciais e não espaciais para explicar ou prever padrões espaciais, sendo preciso enfatizar que nenhuma dessas abordagens permite inferência causal definitiva.

Almeida (2012) afirma que todo processo que se dá no espaço está sujeito a chamada Lei de Tobler (TOBLER, 1970), também conhecida como a Primeira Lei da Geografia, cujo enunciado pode ser estabelecido da seguinte forma: *coisas próximas tendem a ser mais relacionadas do que coisas distantes, tanto no espaço quanto no tempo* e (CRESSIE, 2015) afirma que

a dependência [espacial] *está presente em todas as direções e fica fraca à medida que aumenta a dispersão na localização dos dados.*

Dessa forma, se observações próximas, ou seja, semelhantes em localização, também são semelhantes em valores variáveis, então, o padrão como um todo é tido para exibir autocorrelação espacial positiva (autocorrelação). Por outro lado, diz-se que existe autocorrelação espacial negativa quando as observações, que estão próximas no espaço, tendem a ser mais diferentes em valores variáveis do que as observações que estão mais distantes (em contradição com a lei de Tobler). Já a autocorrelação zero ocorre quando os valores das variáveis são independentes da localização. É importante observar que a autocorrelação espacial invalida a análise estatística convencional e torna a análise de dados espaciais diferente de outras formas de análise de dados (PFEIFFER *et al.*, 2008; FISCHER; WANG, 2011).

Um aspecto crucial da definição de autocorrelação espacial é a determinação de locais próximos, ou seja, aqueles locais em torno de um determinado ponto de dados que podem ser considerados para influenciar a observação naquele ponto de dados. Infelizmente, a determinação dessa vizinhança não ocorre sem algum grau de arbitrariedade (PFEIFFER *et al.*, 2008; FISCHER; WANG, 2011). Formalmente, a associação de observações, na vizinhança, definida para cada local pode ser expressa por meio de uma contiguidade espacial $n \times n$ ou matriz de pesos espaciais W,

$$W = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nn} \end{bmatrix} \tag{2.1}$$

onde n representa o número de localizações (observações). A entrada na linha $i(i = 1,\ldots,n$ e coluna $j(j = 1,\ldots,n$, denotado como W_{ij}, corresponde ao par (i, j) de localizações. Os elementos diagonais da matriz são definidos como zero, por convenção, enquanto os elementos não diagonais $W_{ij}(i \neq j)$ assumem valores diferentes de zero (um, para um binário matriz) quando as localizações i e j são consideradas vizinhas, caso contrário, zero.

As matrizes de pesos espaciais W são uma maneira de representar gráficos em ciência de dados geográficos e estatísticas espaciais. São construções amplamente utilizadas que representam relações geográficas entre as unidades observacionais em um conjunto de dados referenciado espacialmente. Implicitamente, pesos espaciais conectam objetos em uma tabela geográfica usando mutualmente as relações espaciais entre eles. Ao expressar a noção de proximidade geográfica ou conectividade, os pesos espaciais são o principal mecanismo pelo qual as relações espaciais nos dados geográficos são levadas a efeito na análise subsequente (ALMEIDA, 2012; ANDRADE *et al.*, 2007).

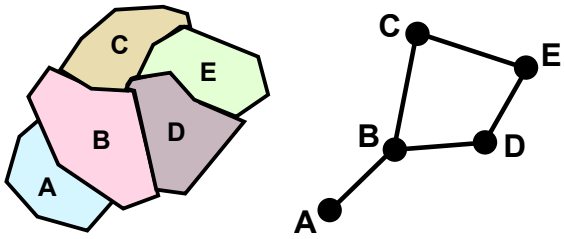
A matriz de pesos espaciais pode ser construída em consonância com a ideia de vizinha baseada na contiguidade, em que duas regiões são vizinhas, caso elas partilhem de uma fronteira

física comum (ALMEIDA, 2012). A ideia é que duas regiões contíguas possuem uma maior interação espacial, dessa forma, podemos definir uma matriz de pesos da seguinte forma:

$$w_{ij} = \begin{cases} 1 & \text{se } i \text{ e } j \quad \text{são contíguos} \\ 0 & \text{se } i \text{ e } j \quad \text{não são contíguos} \end{cases} \tag{2.2}$$

Convencionalmente, é presumido que $w_{ij} = 0$, isto é, a região não é vizinha dela mesma, o que implica que a matriz de contiguidade possua a sua diagonal principal composta por valores nulos. À primeira vista isso parece simples, no entanto, na prática, isso acaba sendo mais complicado. A primeira complicação é que existem diferentes maneiras pelas quais os objetos podem "compartilhar uma borda comum". A Figura 2(a) apresenta um sistema simples de cinco zonas e a Figura 2(b) temos a representação do sistema em grafo. A partir da Equação 2.2 podemos determinar a contiguidade espacial (ou adjacência) (de primeira ordem), considerando $W_{ij} = 1$ se as zonas i e j forem contígua e $W_{ij} = 0$ caso contrário, dessa forma, podemos derivar uma matriz de pesos W na Tabela 1.

Figura 1 – (a) um simples mosaico discreto, (b) representação do mosaico em grafo.



(a) (b)

Fonte: Adaptada de Fischer e Wang (2011).

Tabela 1 – Matriz de pesos espaciais W derivada do sistema da Figura 1: o caso de uma matriz de contiguidade binária de primeira ordem

	A	B	C	D	E
A	0	1	0	0	0
B	1	0	1	1	0
C	0	1	0	0	1
D	0	1	0	0	1
E	0	0	1	1	0

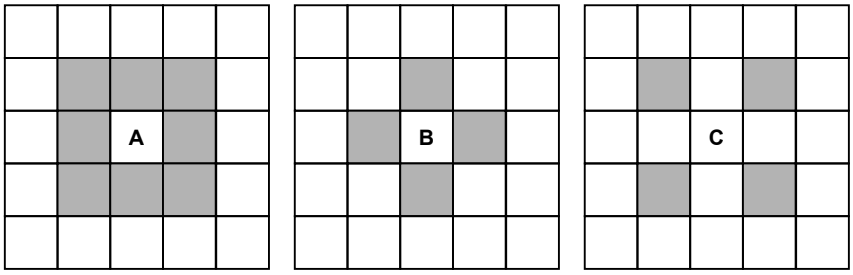
Fonte: Adaptada de Fischer e Wang (2011).

Uma forma comum de expressar as relações de contiguidade/adjacência surge de uma analogia com os movimentos legais que diferentes peças de xadrez podem fazer. A contiguidade é considerada como *torre* (*rook*) caso apenas as fronteiras físicas com a extensão diferente de zero entre as regiões sejam consideradas. Se além das fronteiras com extensão diferente de zero puderem ser considerados os vértices como contíguos, a contiguidade é considerada como *rainha*

(*queen*) se apenas os vértices forem considerados para definir a contiguidade, a convenção é denominada *bispo* (*bishop*) (ALMEIDA, 2012).

As diferentes convenções para a matriz de pesos espaciais são mostradas na Figura 2, em que os vizinhos das regiões A, B e C, respectivamente, estão destacados.

Figura 2 – Convenção de contiguidade: (a) rainha, (b) torre e (c) bisbo.



(a) (b) (c)

Fonte: Adaptada de Almeida (2012).

Outra forma de critério de proximidade na definição dos pesos espaciais é a distância geográfica. A ideia por trás é que duas regiões mais próximas geograficamente têm uma maior interação espacial. Uma matriz W muito adotada na literatura para esta situação é a matriz dos k vizinhos mais próximos (do inglês, *K-Nearest Neighbors (KNN)*) (ALMEIDA, 2012; FISCHER; WANG, 2011).

2.3.1 Visualização de Dados Espaciais

A forma mais simples e intuitiva de análise exploratória é a visualização de valores extremos nos mapas (CÂMARA *et al.*, 2004). O mapa é o meio mais estabelecido e convencional de exibir dados de área. Há uma variedade de maneiras de atribuir dados de variáveis contínuas a determinadas unidades de área que são predefinidas. Na prática, porém, nenhum é isento de problemas. Vale ressaltar que o uso de diferentes pontos de corte da variável, induz a visualização de diferentes aspectos(CÂMARA *et al.*, 2004; FISCHER; WANG, 2011).

Salienta-se que os mapas coropléticos são a forma de exibição mais utilizada para representação de áreas (FISCHER; WANG, 2011). Trata-se de um mapa onde cada uma das áreas é colorida ou sombreada de acordo com uma escala discreta baseada no valor da variável (atributo) de interesse dentro daquela área. O número de classes (categorias) e os intervalos de classes (categorias) correspondentes podem ser baseados em vários critérios diferentes.

Dessa forma, os mapas coropléticos desempenham um papel proeminente na ciência de dados geográficos, pois permitem aos pesquisadores exibir atributos ou variáveis não geográficas em um mapa geográfico e é a forma usual de apresentação de dados agregados por áreas (CÂMARA *et al.*, 2004).

Todavia, a eficácia de um mapa coroplético depende na maioria do propósito do mapa. Qual mensagem você deseja comunicar moldará quais opções são preferíveis em relação a outras. Podem-se considerar três dimensões sobre as quais colocar o pensamento intencional valerá a pena. Os mapas coropléticos giram em torno de: primeiro, selecionar um número de grupos menor que n em que todos os valores no conjunto de dados serão mapeados; segundo, identificar um algoritmo de classificação que execute tal mapeamento, seguindo algum princípio alinhado ao interesse da equipe; e terceiro, uma vez que se sabe em quantos grupos reduzir–se-ão todos os valores nos dados, qual cor é atribuída a cada grupo para garantir que ele codifique as informações que se quer refletir. Em termos gerais, o esquema de classificação define o número de classes, bem como as regras de atribuição; enquanto uma boa simbolização transmite informações sobre a diferenciação de valor entre as classes (FISCHER; WANG, 2011; CÂMARA *et al.*, 2004).

2.3.1.1 Classificação de dados quantitativos

Selecionar o número de grupos aos quais se quer atribuir os valores nos dados da equipe e como cada valor é atribuído a um grupo, pode ser visto como um problema de classificação (LONGLEY *et al.*, 2005; FISCHER; WANG, 2011). A classificação de dados considera o problema de particionar os valores dos atributos em grupos mutuamente exclusivos e exaustivos. A forma precisa como isso será feito, terá como condição a função da escala de mensuração do atributo em questão. Para atributos quantitativos (escalas ordinais, intervalares, de razão), as classes terão uma ordenação explícita. Mais formalmente, o problema de classificação é definir limites de classe de modo que:

$$c_j < u_i \leq c_{i+1} \forall y_i \in C_j \tag{2.3}$$

em que y_i é o valor do atributo para localização espacial i, j é um índice de classe, e c_j representa o limite inferior do intervalo j. Diferentes esquemas de classificação são obtidos com base em sua definição dos limites de classe. A escolha do esquema de classificação deve levar em consideração a distribuição estatística dos valores de atributo, bem como, o objetivo do nosso mapa (por exemplo, destacar valores discrepantes versus descrever com precisão a distribuição de valores) (LONGLEY *et al.*, 2005; FISCHER; WANG, 2011; LOPES *et al.*, 2022).

Em Longley *et al.* (2005) é sugerida quatro esquemas básicos de classificação que podem ser utilizadas para dividir dados de área de intervalo e razão em categorias: (i) Intervalos iguais, (ii) Quantis, (iii) Desvio médio-padrão e (iv) *Natural breaks*, definidas a seguir.

Intervalos iguais

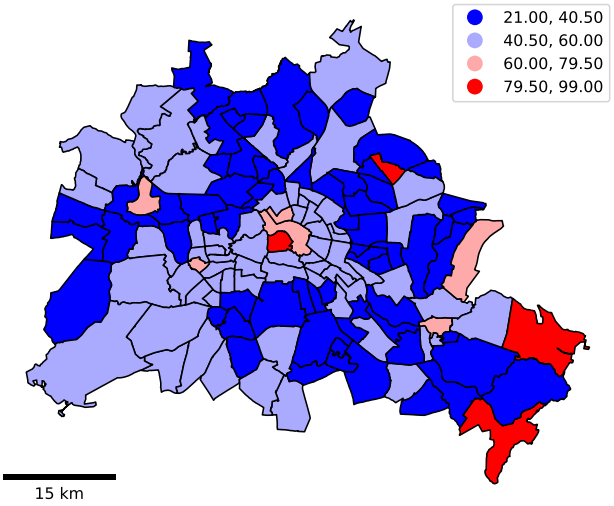
A abordagem de Freedman-Diaconis (FREEDMAN; DIACONIS, 1981) fornece uma regra para determinar a largura e, por sua vez, o número de caixas para a classificação. Esse é um caso especial de um classificador mais geral conhecido como “intervalos iguais”, no qual cada uma das caixas tem a mesma largura no espaço de valor. Para um determinado valor de

k, a classificação de intervalos iguais divide o intervalo do espaço de atributo em k intervalos de comprimento igual, sendo que cada intervalo tem uma largura $w = \frac{x_0 - x_{n-1}}{k}$. Assim, a classe máxima é $(x_{n-1} - w, x_{n-1}]$ e a primeira classe é $(-\infty, x_{n-1} - (k-1)w]$.

Intervalos iguais têm as vantagens duplas de simplicidade e facilidade de interpretação. No entanto, essa regra considera apenas os valores extremos da distribuição e, em alguns casos, isso pode resultar em uma ou mais classes esparsas (LONGLEY *et al.*, 2005; FISCHER; WANG, 2011; LOPES *et al.*, 2022).

Para exemplificar a utilização desse classificador, a Figura 3 apresenta um mapa coroplético da distribuição dos preços médios dos imóveis nos distritos de Berlim, capital da Alemanha, com o esquema de intervalos iguais.

Figura 3 – Distribuição do preço médio dos imóveis nos dos preços médios dos imóveis nos distritos de Berlim pela abordagem de intervalos iguais.



Fonte: Adaptada de Longley *et al.* (2005).

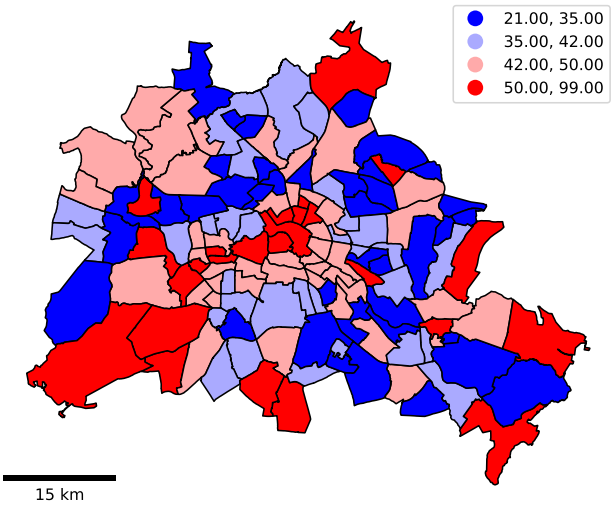
Observe na Figura 3 que cada um dos intervalos tem igual largura de $39.5. Deve-se notar também que a primeira classe é fechada no limite inferior, em contraste com a abordagem geral definida acima.

Quantis

Para evitar o problema potencial de classes esparsas, os *quantis* da distribuição podem ser usados para identificar os limites das classes. De fato, cada classe terá aproximadamente $\left|\frac{n}{k}\right|$ observações usando o classificador de quantil. Se $k = 5$ os quintis da amostra são usados para definir os limites superiores de cada classe, como pode ser observado na Figura 4.

Observe que na Figura 4, embora os números de valores, em cada classe, sejam aproximadamente iguais, as larguras dos quatro primeiros intervalos são bastante diferentes. Embora os *quantis* evitem a armadilha das classes esparsas, essa classificação não está livre de problemas,

Figura 4 – Distribuição do preço médio dos imóveis nos dos preços médios dos imóveis nos distritos de Berlim pela abordagem de *quantis*.



Fonte: Adaptada de Longley *et al.* (2005).

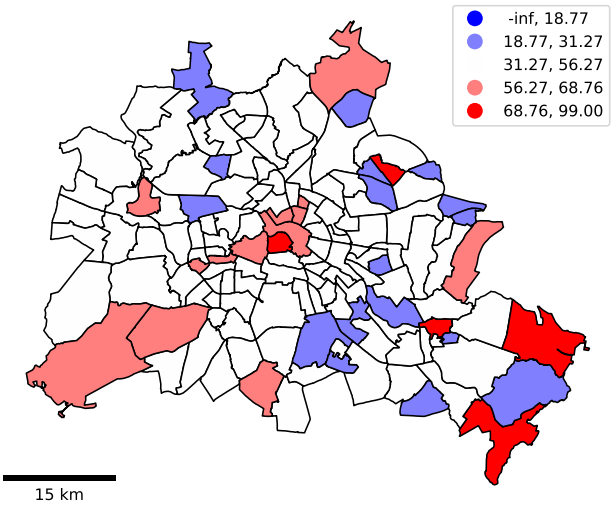
pois as larguras variáveis dos intervalos podem ser marcadamente diferentes, o que pode levar a problemas de interpretação. Um segundo desafio enfrentado pelos *quantis* surge quando há inúmeros valores duplicados na distribuição, de modo que os limites para uma ou mais classes se tornam ambíguos. Por exemplo, se alguém tivesse uma variável com $n = 20$ mas 10 das observações assumiram o mesmo valor que foi o mínimo observado, então para valores de $k > 2$, os limites de classe tornam-se mal definidos, visto que uma simples regra de divisão n/k no valor observado classificado dependeria de como os empates são tratados na classificação.

Desvio médio-padrão

O terceiro modo de classificação utiliza a média amostral $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e desvio padrão amostral $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})}$ para definir limites das classes como alguma distância da média da amostra, com a distância sendo um múltiplo do desvio padrão. Por exemplo, uma definição comum para $k = 5$ é definir o limite superior da primeira classe para dois desvios padrão $(c_0^u = \bar{x} - 2s)$, e as classes intermediárias tenham limites superiores dentro de um desvio padrão $(c_1^u = \bar{x} - s, c_2^u = \bar{x} + s, c_3^u = \bar{x} + 2s)$. Quaisquer valores maiores (menores) do que dois desvios padrão acima (abaixo) da média são colocados na classe superior (inferior).

Esse classificador é melhor usado quando os dados são normalmente distribuídos ou, pelo menos, quando a média da amostra é uma medida significativa para ancorar a classificação. Claramente, esse não é o caso de nossos dados de renda, pois a inclinação positiva resulta em perda de informações quando usamos o desvio padrão. A falta de simetria leva a um limite superior inadmissível para a primeira classe, bem como, a uma concentração da grande maioria dos valores na classe média. A Figura 5 apresenta um exemplo de utilização desse modo de

Figura 5 – Distribuição do preço médio dos imóveis nos dos preços médios dos imóveis nos distritos de Berlim pela abordagem de desvio médio-padrão.



Fonte: Adaptada de Longley *et al.* (2005).

definir as classes.

Quebra Natural (Natural breaks)

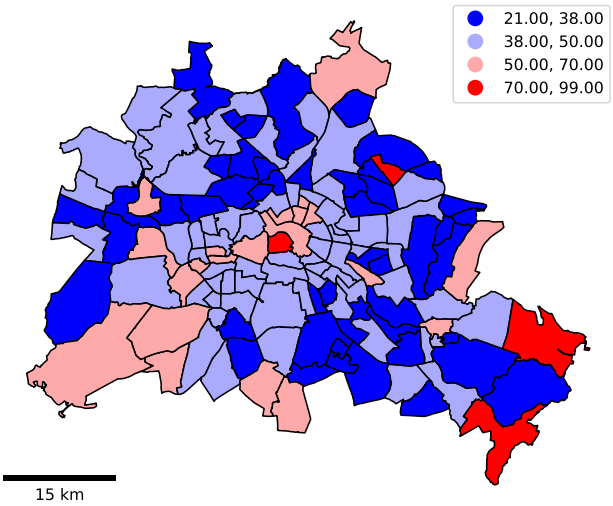
É importante, dessa forma, que se conceitualize o que seriam as classes. As classes são, portanto, definidas de acordo com alguns agrupamentos naturais dos valores de dados. As quebras (*breaks* podem ser impostas com base em pontos de quebra conhecidos por serem relevantes em um determinado contexto de aplicação, como frações e múltiplos de níveis médios de renda, ou limiares pluviométricos da vegetação (‘árido’, ‘semi-árido’, ‘ temperado’, etc.). Essa é uma atribuição dedutiva de quebras, enquanto classificações indutivas de valores de dados podem ser realizadas usando ferramentas GIS para procurar saltos relativamente grandes em valores de dados. A Figura 6 apresenta um exemplo de utilização desse modo de definir as classes.

É possível também adotar outras formas de classificação, como *Maximum breaks*, *Jenks Caspall* e *Fisher Jenks* (LOPES *et al.*, 2022).

Quebra Máxima (Maximum breaks)

O classificador *maximum breaks* determina a posição dos pontos de quebra entre as classes ao considerar a diferença entre os valores ordenados. Em vez de analisar individualmente o valor de cada dado do conjunto, ele observa a distância entre cada valor e o próximo na sequência ordenada. O classificador, assim, coloca os $k-1$ pontos de quebra entre os pares de valores mais distantes em toda a sequência e segue uma ordem decrescente com base no tamanho das diferenças.

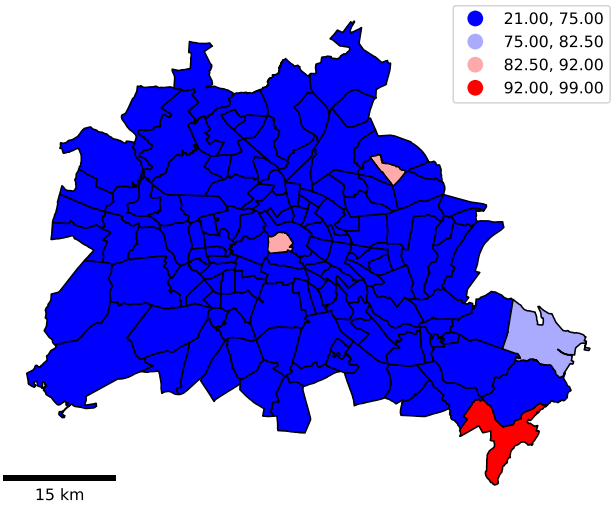
Figura 6 – Distribuição do preço médio dos imóveis nos dos preços médios dos imóveis nos distritos de Berlim pela abordagem *natural breaks*.



Fonte: Adaptada de Longley *et al.* (2005).

A Figura 7 apresenta um exemplo de utilização desse modo de definir as classes.

Figura 7 – Distribuição do preço médio dos imóveis nos dos preços médios dos imóveis nos distritos de Berlim pela abordagem *maximum breaks*.



Fonte: Adaptada de Longley *et al.* (2005).

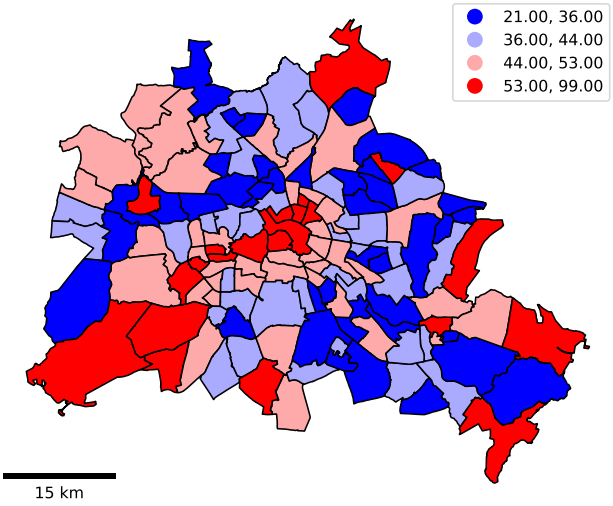
Jenks Caspall

Originalmente proposto por Jenks e Caspall (1971), aborda o desafio da classificação de uma perspectiva heurística, opondo-se à determinística. Dessarte visa minimizar a soma dos desvios absolutos em torno das médias das classes. A abordagem começa com um número pré-especificado de classes e um conjunto inicial arbitrário de quebras de classe - por exemplo,

usando *quintis*. O algoritmo tenta melhorar a função objetivo considerando o movimento de observações entre classes adjacentes. Por exemplo, o maior valor no *quintil* mais baixo seria considerado para o movimento para o segundo quintil, enquanto o valor mais baixo, no segundo, *quintil* seria considerado para um possível movimento para o primeiro *quintil*. O movimento candidato que resultar na maior redução na função objetivo seria feito e o processo continua até que nenhum outro movimento de melhoria seja possível. O algoritmo *Jenks Caspall* é o caso unidimensional do algoritmo *K-Means* amplamente utilizado para agrupamento.

Na Figura 8 temos a aplicação dessa abordagem para o conjunto de dados dos preços médios dos imóveis por distritos em Berlim.

Figura 8 – Distribuição do preço médio dos imóveis nos dos preços médios dos imóveis nos distritos de Berlim pela abordagem pela abordagem *Jenks Caspall*.



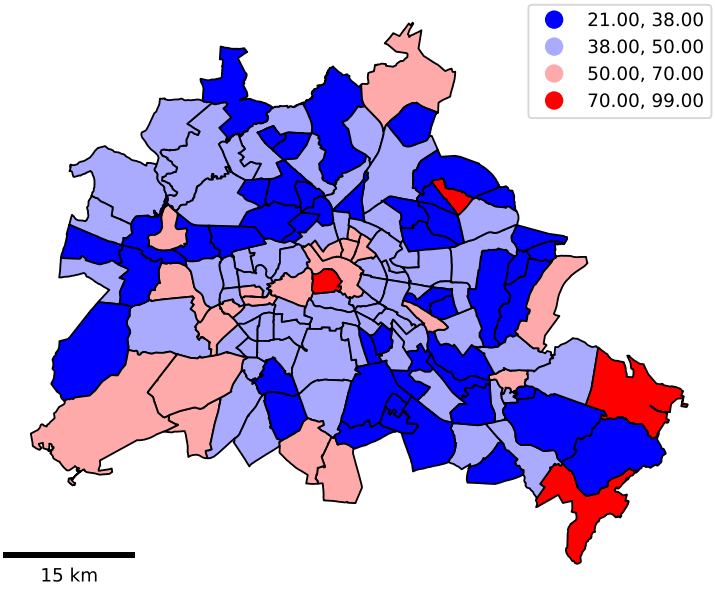
Fonte: Adaptada de Longley *et al.* (2005).

Fisher Jenks

O segundo algoritmo ótimo adota uma abordagem de programação dinâmica para minimizar a soma dos desvios absolutos em torno das medianas de classe. Em contraste com a abordagem *Jenks Caspall*, o *Fisher Jenks* garante a produção de uma classificação ideal para um número pré-especificado de classes. Na Figura 9, temos a aplicação dessa abordagem para o conjunto de dados dos preços médios dos imóveis por distritos em Berlim.

Como um caso especial de agrupamento, a definição do número de classes e os limites de classe representam um problema para o projetista do mapa. Para a classificação de mapas, um critério de otimização comum é uma medida de ajuste. Uma métrica possível de utilização é o "desvio absoluto em torno das medianas de classe" (*Absolute Deviation Around Class Medians (ADCM)*) (LOPES *et al.*, 2022). O ADCM fornece uma medida de ajuste que permite a comparação de esquemas de classificação para o mesmo valor de k e não é sensível à presença de

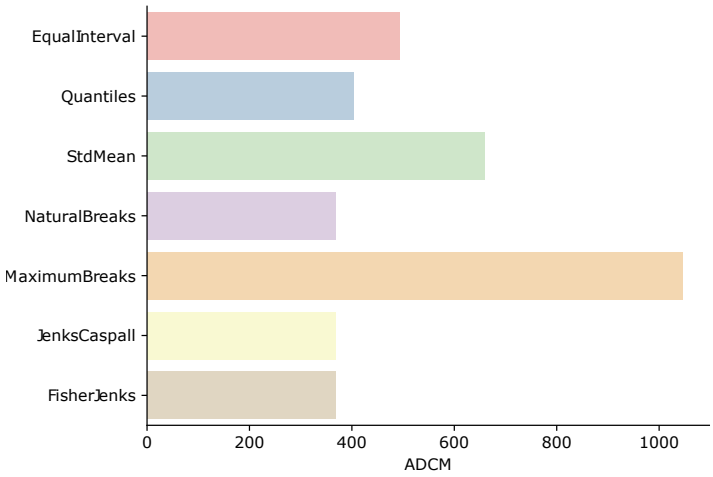
Figura 9 – Distribuição do preço médio dos imóveis nos dos preços médios dos imóveis nos distritos de Berlim pela abordagem *Fisher Jenks*.



Fonte: Adaptada de Longley *et al.* (2005).

valores discrepantes (LONGLEY *et al.*, 2005; FISCHER; WANG, 2011; LOPES *et al.*, 2022). O ADCM nos dá uma noção de quão "compacto" é cada grupo. A Figura 10 mostra o ADCM para diferentes classificadores sobre os dados dos preços médios dos imóveis na cidade de Berlim, considerando $k = 5$.

Figura 10 – ADCM dos classificadores usados na base de dados dos preços médios dos imóveis na cidade de Berlim.



Fonte: Elaborada pelo autor.

É possível observar na Figura 10 que o classificador *Jenks Caspall*, *Natural Breaks* e

Fisher Jenks dominam todos os outros classificadores para $k = 5$ com um ADCM de 368,17; 368,83 e 368,83, respectivamente. Vale ressaltar que quanto menor melhor, já que produz classes mais compactas.

2.3.2 Medidas Globais e Testes para Autocorrelação Espacial

Salienta-se que o foco da análise exploratória de dados espaciais está em medir e exibir padrões globais e locais de associação espacial, os indicam não-estacionariedade local, descobrindo ilhas de heterogeneidade espacial, e assim por diante (FISCHER; WANG, 2011). Uma vez determinada a matriz de pesos W do atributo do fenômeno em estudo é necessário usar alguma estatística de teste que averígue a aleatoriedade da distribuição espacial da variável sob estudo formal global (ALMEIDA, 2012).

Sendo que a aleatoriedade espacial significa que os valores de um atributo numa região não dependem dos valores deste atributo nas regiões vizinhas. Uma forma de verificar essa relação de dependência é por meio de uma medida de autocorrelação espacial. As medidas de autocorrelação espacial tratam da covariação ou correlação entre observações vizinhas de uma variável. Desta maneira, podem-se comparar dois tipos de informação: similaridade de observações (semelhança de valor) e similaridade entre locais (FISCHER; WANG, 2011; LONGLEY *et al.*, 2005; ALMEIDA, 2012; LOPES *et al.*, 2022)

A noção de autocorrelação espacial refere-se à existência de uma "relação funcional entre o que acontece em um ponto do espaço e o que acontece em outro" (ANSELIN, 1988). A autocorrelação espacial, portanto, tem a ver com o grau em que a similaridade de valores entre observações em um conjunto de dados está relacionada à similaridade nas localizações de tais observações. Isso é semelhante à ideia tradicional de correlação entre duas variáveis que informa sobre como os valores de uma variável mudam em função dos da outra, embora, com algumas diferenças. De maneira semelhante, a autocorrelação espacial também está relacionada (mas distinta) à contraparte temporal, à autocorrelação temporal, que relaciona o valor de uma variável em um determinado momento com os de períodos anteriores. Em contraste com essas outras ideias de correlação, a autocorrelação espacial relaciona o valor da variável de interesse em um determinado local com valores da mesma variável em outros locais. Uma forma alternativa de entender o conceito é como o grau de informação contido no valor de uma variável, em um determinado local, sobre o valor dessa mesma variável, em outros locais, (ALMEIDA, 2012; ANDRADE *et al.*, 2007; ANSELIN; FLORAX; REY, 2013; ANSELIN, 2005).

Para entender melhor a noção de autocorrelação espacial, é útil começar considerando como é o mundo na sua ausência. Uma ideia chave, nesse contexto, é a da aleatoriedade espacial: uma situação onde a localização de uma observação não fornece nenhuma informação sobre o seu valor. Em outras palavras, uma variável é espacialmente aleatória se sua distribuição não segue um padrão espacial discernível. A autocorrelação espacial pode assim ser definida como a "ausência de aleatoriedade espacial" (ALMEIDA, 2012; ANDRADE *et al.*, 2007; ANSELIN;

FLORAX; REY, 2013; ANSELIN, 2005).

Medidas e testes de autocorrelação espacial (associação) podem ser diferenciados pelo escopo ou escala de análise. Geralmente, distinguem-se entre medidas globais e locais. Global implica que todos os elementos, na matriz W, são usados para avaliar a autocorrelação espacial. Ou seja, todas as associações espaciais de áreas são incluídas no cálculo da autocorrelação espacial. Isso rende um valor para autocorrelação espacial para qualquer matriz de pesos espaciais. Em contraste, as medidas locais são focadas, ou seja, elas avaliam a autocorrelação espacial associada a uma ou algumas unidades de área específicas (FISCHER; WANG, 2011; ALMEIDA, 2012).

Medidas globais de autocorrelação espacial comparam o conjunto de similaridade de valor (observação) M_{ij} com o conjunto de similaridade espacial W_{ij}, combinando-os em um único índice de um produto cruzado, isto é:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}W_{ij} \tag{2.4}$$

no qual n é o número de áreas na amostra, i, j quaisquer duas das unidades de área, W_{ij} a similaridade das localizações de i e j, com $W_{ii} = 0$ para todo i e M_{ij} a similaridade das observações de i e j da variável do fenômeno em estudo (FISCHER; WANG, 2011).

Várias maneiras têm sido sugeridas para medir a similaridade de valor (associação) M_{ij} dependendo da escala da variável. Para variáveis nominais, a abordagem é definir M_{ij} como um se i e j tiverem o mesmo valor de variável e zero caso contrário. Para variáveis ordinais, a similaridade de valor é, geralmente, baseada na comparação das classificações de i e j. Para variáveis de intervalo, tanto a diferença quadrada $(z_i - z_j)^2$, com z sendo o valor (observação) da variável de interesse para a região i, e quanto o produto $(z_i - \bar{x})(z_j - \bar{x})$ são comumente usados, em que $\bar{z}$ denota a média dos valores z (FISCHER; WANG, 2011; ALMEIDA, 2012).

As duas medidas que têm sido mais amplamente utilizadas para o caso de unidades de área e variáveis de escala de intervalo são as estatísticas I de Moran e o índice c de Geary (FISCHER; WANG, 2011; ALMEIDA, 2012). Ambas indicam o grau de associação espacial conforme resumido para todo o conjunto de dados. O I de Moran usa produtos cruzados para medir a associação e c de Geary diferenças ao quadrado (ALMEIDA, 2012).

Estatística I de Moran

Moran (1948) propôs a elaboração de um coeficiente de autocorrelação espacial, em que usa a média de autocovariância na forma de produto cruzado. Assim surgia o primeiro coeficiente de autocorrelação espacial, denominado de I de Moran, definido pela Equação 2.5.

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2} \tag{2.5}$$

em que n é o número de observações, z_i é o valor padronizado da variável de interesse na região i, e w_{ij} é a célula correspondente à i-ésima linha da j-ésima coluna da matriz de pesos espaciais.

A estatística I de Moran é a estatística mais difundida e constitui em uma espécie de coeficiente de autocorrelação, ou seja, é a relação da autocovariância do tipo produto cruzado pela variância dos dados (ANDRADE *et al.*, 2007; ALMEIDA, 2012).

Uma abordagem alternativa para entender a intuição por trás de sua matemática, é através de uma interpretação gráfica e essa interpretação gráfica pode ser efetuada através do diagrama de dispersão de Moran, também conhecido como Moran *Plot*. O Moran *Plot* é uma maneira de visualizar um conjunto de dados espaciais para explorar a natureza e a força da autocorrelação espacial. É essencialmente um gráfico de dispersão tradicional em que a variável de interesse é exibida em relação à defasagem espacial da variável de interesse. Para poder interpretar valores como acima ou abaixo da média, a variável de interesse é geralmente padronizada subtraindo sua média (ANDRADE *et al.*, 2007; ALMEIDA, 2012).

Índice C de Geary

A razão de contiguidade proposta por Geary (1954), é dado pela Equação 2.6.

$$C = \frac{(n-1)}{2\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - y_j)^2}{\sum_i (y_i - \bar{y})^2} \tag{2.6}$$

no qual n é o número de observações, w_{ij} é a célula em uma matriz binária W expressando se i e j são vizinhos ($w_{ij} = 1$) ou não ($w_{ij} = 0$), y_i é a i-ésima observação da variável de interesse, e $\bar{y}$ é sua média amostral. Quando comparado à estatística I de Moran, fica evidente que ambas as medidas comparam a relação de Y dentro da vizinhança local de cada observação para aquela em toda a amostra. No entanto, também existem diferenças sutis. Enquanto a estatística I de Moran usa produtos cruzados nos valores padronizados, o índice C de Geary usa diferenças nos valores sem qualquer padronização, sendo computacionalmente mais exigente (FISCHER; WANG, 2011; ALMEIDA, 2012).

Os testes de autocorrelação espacial são regras de decisão baseadas em estatísticas como a estatística I de Moran e índice c Geary para avaliar até que ponto o arranjo espacial observado de valores de dados se afasta da hipótese nula de que o espaço não importa. Essa hipótese implica que as áreas próximas não afetam umas às outras de forma que haja independência e aleatoriedade espacial (FISCHER; WANG, 2011).

Em contraste, sob a hipótese alternativa de autocorrelação espacial (associação espacial, dependência espacial), o interesse recai sobre os casos em que grandes valores são cercados por outros grandes valores em áreas próximas, ou pequenos valores são cercados por grandes valores e vice-versa. O primeiro é referido como autocorrelação espacial positiva e o último, como autocorrelação espacial negativa. A autocorrelação espacial positiva implica num agrupamento

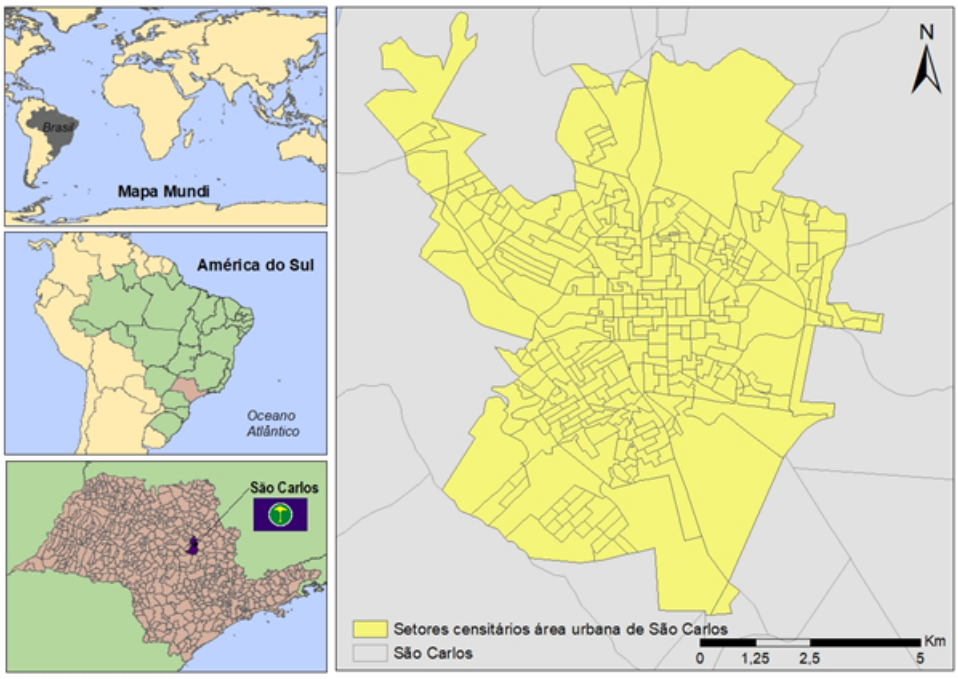
espacial de valores semelhantes, enquanto a autocorrelação espacial negativa implica num padrão quadriculado de valores (FISCHER; WANG, 2011; ALMEIDA, 2012).

A autocorrelação espacial é considerada presente quando a estatística de autocorrelação espacial calculada para um determinado padrão assume um valor maior, em comparação com o que seria esperado sob a hipótese nula de nenhuma associação espacial. O que é visto como significativamente maior depende da distribuição da estatística de teste (FISCHER; WANG, 2011; ALMEIDA, 2012).

Em princípio, existem duas abordagens principais para testar os I valores observados quanto ao afastamento significativo da hipótese de autocorrelação espacial zero (MORAN, 1948; FISCHER; WANG, 2011). O primeiro é o teste de permutação aleatória. Sob a hipótese de randomização, o valor observado de I é avaliado em relação ao conjunto de todos os valores possíveis que podem ser obtidos por permuta aleatória das observações sobre os locais no conjunto de dados.

Para ilustrar a noção de autocorrelação espacial e suas diferentes variantes, utilizar-se-á as notificações confirmadas de casos de dengue por setor censitário [1] da zona urbana do município de São Carlos (Figura 11), localizado na região central do estado de São Paulo, disponibilizados pela vigilância epidemiológica do município.

Figura 11 – Município de São Carlos-SP e seu perímetro urbano.

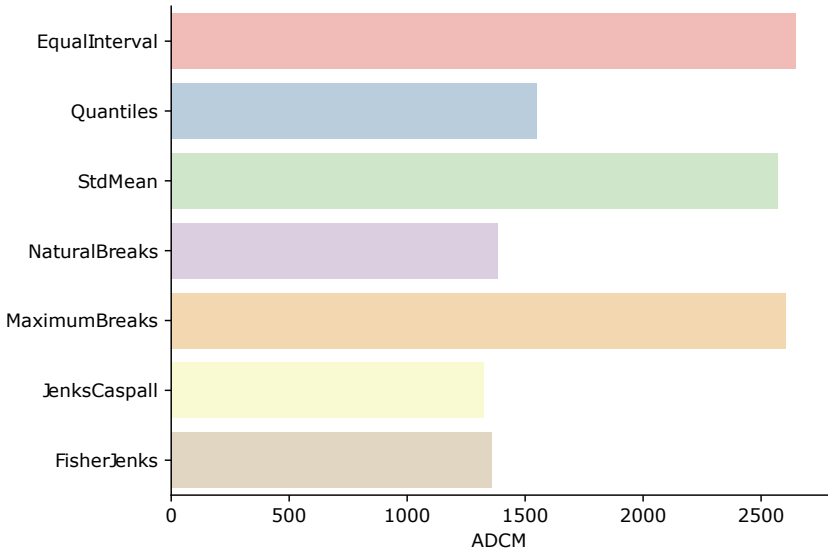


Fonte: Lopes *et al.* (2023b).

[1] O setor censitário é a unidade territorial estabelecida para fins de controle cadastral, formado por área contínua, situada em um único quadro urbano ou rural com dimensão e número de domicílios que permitam o levantamento por um recenseador (CENSO, 2010).

Para gerar essa base de dados de casos de dengue, foram considerados todos os casos confirmados de dengue notificados no Sistema de Informação de Agravos de Notificação - Sinan Dengue/Chikungunya dos residentes da área urbana do município de São Carlos-SP no período de 1 de janeiro à 31 de dezembro do ano de 2019. Essas notificações foram georreferenciadas e agregadas aos setores censitários por meio de uma função de *join spatial*. Visando encontrar número de classes e os limites de cada classe, foi calculado o ADCM dos casos de dengue, como pode ser observado na Figura 12, o menor ADCM é o classificador *Jenks Caspall*.

Figura 12 – ADCMs dos casos confirmados de dengue do município de São Carlos-SP.



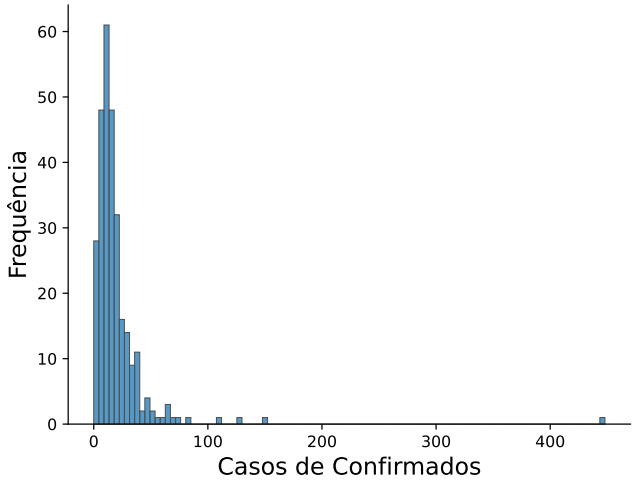
Fonte: Elaborada pelo autor.

Uma vez determinado o classificador que melhor descreve o conjunto de dados, pode-se analisar a distribuição do atributo em estudo. A Figura 13 mostra o histograma com a frequência de casos confirmados de dengue nos setores censitários. Já a Figura 14 apresenta a representação coroplética dos casos confirmados de dengue por setor censitário, classificados segundo o algoritmo de *Jenks Caspall*.

Ao analisar um mapa, como o da Figura 14, buscam-se padrões do ponto de vista espacial e um dos aspectos necessários que temos que responder, como já mencionado, é verificar se o evento em estudo e os fatores relacionados a ele possuem distribuição espacialmente condicionada (dependência espacial), ou seja, se os valores do atributo em estudo, em uma determinada região, dependem ou não dos valores desse atributo nas regiões vizinhas, tarefa extremamente difícil de se realizar visualmente (ALMEIDA, 2012).

A Figura 15 mostra a relação entre a porcentagem padronizada casos confirmados de dengue e sua defasagem espacial que, devido à padronização por linha de w, pode ser interpretado como a densidade padronizada média dos casos confirmados de dengue na vizinhança de cada

Figura 13 – Histograma dos casos confirmados de dengue por setor censitário do município de São Carlos-SP.



Fonte: Elaborada pelo autor.

observação. A fim de orientar a interpretação do gráfico, um ajuste linear também é incluído. Essa linha representa o melhor ajuste linear ao gráfico de dispersão ou, em outras palavras, qual é a melhor forma de representar a relação entre as duas variáveis como uma linha reta.

O gráfico da Figura 15 mostra uma relação positiva entre ambas as variáveis. Isso indica a presença de autocorrelação espacial positiva: valores semelhantes tendem a se localizar próximos uns dos outros. Isso significa que a tendência geral é que os valores altos estejam próximos de outros valores altos e que os valores baixos sejam cercados por outros valores baixos. Isso, no entanto, não significa que este padrãoseja o único caso no conjunto de dados: é claro que pode haver situações particulares em que valores altos são cercados por valores baixos e vice-versa. Mas isso significa que, se tivéssemos que resumir o padrão principal dos dados em termos de quão agrupados são os valores semelhantes, a melhor maneira seria dizer que eles são positivamente correlacionados, portanto, agrupados no espaço.

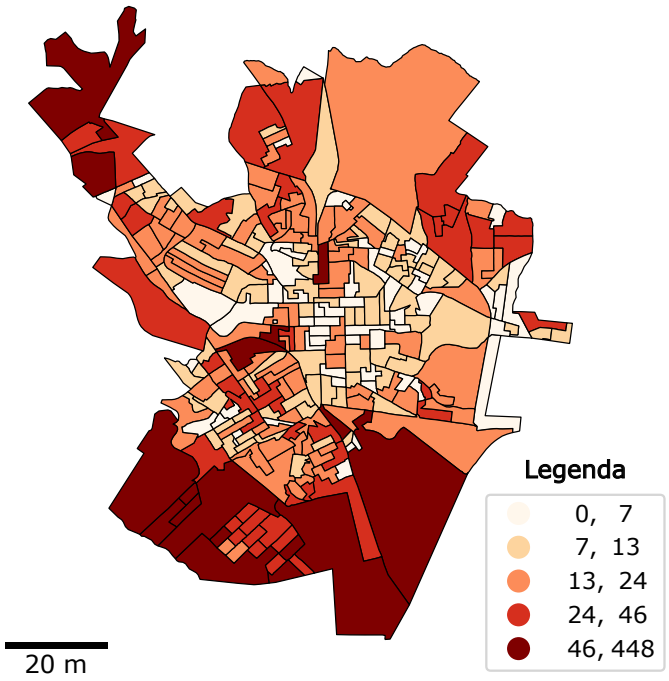
Outro índice global

A estatística *I* de Moran é provavelmente a mais utilizada para autocorrelação espacial global, porém não é a única. Apresentam-se duas medidas adicionais comuns no trabalho aplicado. Embora todos considerem a autocorrelação espacial, diferem na forma como o conceito é abordado na especificação de cada teste.

Índice G de Getis e Ord

Originalmente proposto por Getis e Ord (2010), o índice *G* é uma medida de autocorrelação espacial de natureza global. Na primeira versão dessa estatística, ela é computada para

Figura 14 – Representação coroplética da distribuição dos casos confirmados de dengue por setor censitário em 2019 do município de São Carlos-SP.



Fonte: Elaborada pelo autor.

apenas valores positivos de uma variável por definir um conjunto de vizinhos para cada região como aquelas observações que fiquem dentre de uma distância de corte (*cutt-off*) fixa (d) da região (ALMEIDA, 2012). O índice G de Getis e Ord é definida pela Equação 2.7.

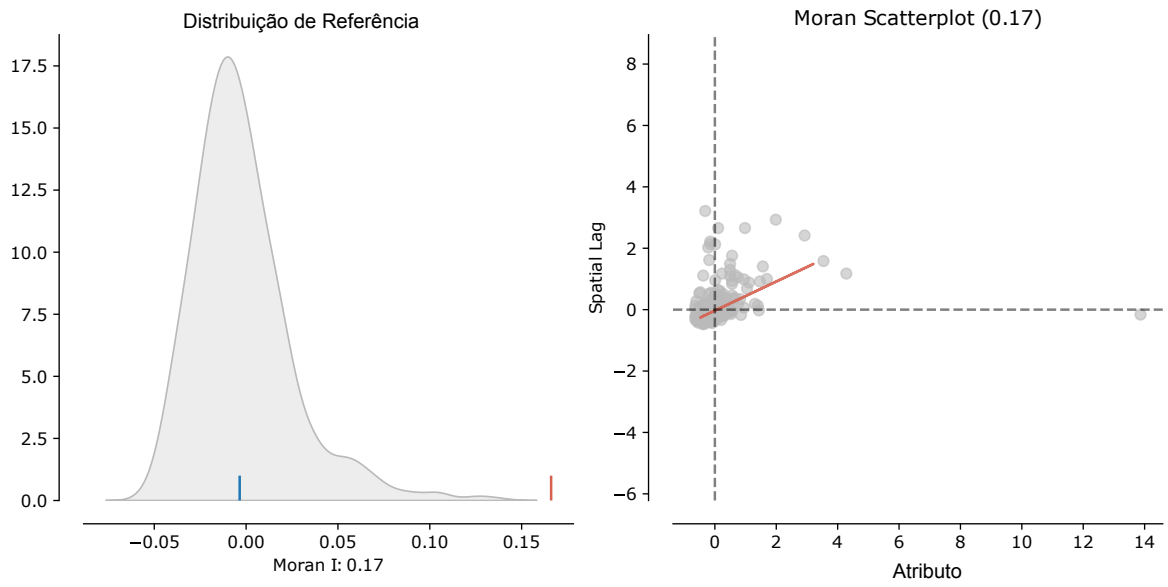
$$G(d) = \frac{\sum_i \sum_j w_{ij}(d) y_i y_j}{\sum_i \sum_j y_i y_j} \tag{2.7}$$

onde w_{ij} é o peso binário atribuído na relação entre as observações i e j seguindo um critério de distância.

2.3.3 Medidas e testes locais para autocorrelação espacial

As estatísticas globais de autocorrelação espaciais, vistas na Seção anterior, fornecem padrões de associação linear espacial, ou seja, o grau em que o conjunto dos dados está agrupado, disperso ou distribuído aleatoriamente (ALMEIDA, 2012; FOTHERINGHAM; BRUNSDON; CHARLTON, 2000). A presença de autocorrelação espacial tem implicações importantes para análises estatísticas subsequentes. De uma perspectiva substantiva, a autocorrelação espacial poderia refletir a operação de processos que geram associação entre os valores em localidades próximas. Isso pode representar transbordamentos, onde os resultados, em um local, influenciam outros locais ou pode indicar contágio, onde os resultados, em um local, influenciam causalmente outros locais.

Figura 15 – Gráfico de dispersão de Moran.



Fonte: Elaborada pelo autor.

Apesar de sua importância, as medidas globais de autocorrelação espacial são estatísticas de "mapa inteiro". Elas fornecem um único resumo para um conjunto de dados inteiro, isto é, as medidas podem dizer se os valores no mapa se agrupam (ou se dispersam) em geral, mas não informará sobre onde estão os agrupamentos específicos (ou valores discrepantes) (ALMEIDA, 2012; FOTHERINGHAM; BRUNSDON; CHARLTON, 2000).

I de Moran local

Proposto por Anselin (1995), os Indicadores Locais de Associação Espacial (do inglês, *Local Indicators of Spatial Association (LISA)*), visa identificar casos em que o valor de uma observação e a média de seus arredores são mais semelhantes (alto-alto, do inglês *high-high* - HH ou baixo-baixo, do inglês *low-low* - LL no gráfico de dispersão de Moran) ou diferentes (alto-baixo, do inglês *high-low* - HL ou baixo-alto, do inglês *low-high* - LH) do que esperar-se-ia do puro acaso. O mecanismo para fazer isso é semelhante ao do índice I de Moran global, mas aplicado nesse caso a cada observação. Isso resulta em tantas estatísticas quanto as observações originais. A representação formal da estatística pode ser escrita pela Equação 2.8.

$$I_i = \frac{z_i}{m_2} \sum_j w_{ij} z_j ; m_2 = \frac{\sum_i z_i^2}{n} \tag{2.8}$$

onde m_2 é o segundo momento (variância) da distribuição de valores nos dados, $z_i = u_i - \bar{y}, w_{ij}$, é o peso espacial para o par de observações i e j, e n é o número de observações.

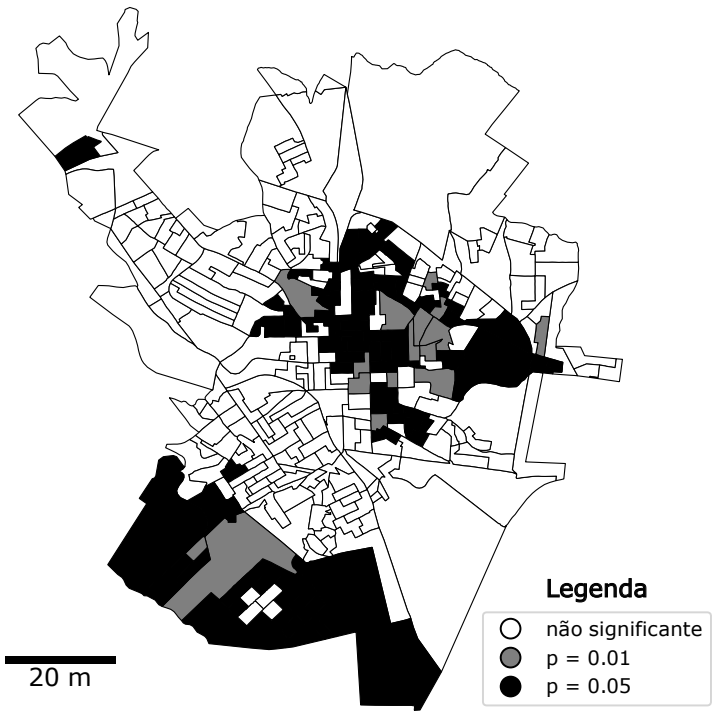
LISA é amplamente utilizada, em muitos campos, para identificar agrupamentos geográ-

ficos de valores ou encontrar discrepâncias geográficas. É uma ferramenta útil que pode retornar rapidamente as áreas em que os valores estão concentrados e fornecer evidências sugestivas sobre os processos que podem estar em ação. Por esses motivos, eles têm um lugar privilegiado na caixa de ferramentas da ciência de dados geográficos. Entre muitas outras aplicações, LISA têm sido usadas para identificar aglomerados geográficos de pobreza (DAWSON *et al.*, 2018), delinear áreas de atividade econômica particularmente alta/baixa (TORRES-PRECIADO; POLANCO-GAYTAN; TINOCO-ZERMEÑO, 2014) ou identificar aglomerados de doenças contagiosas (ZHANG *et al.*, 2020).

Como já visto, pode-se avaliar a associação linear espacial localizada pelo I_i de Moran local. Para cada observação é computado um I_i. Assim, obtemos n computações da estatística I_i e os seus respectivos níveis de significância. Essa copiosa quantidade de informações pode confundir, se colocada em tabelas (ALMEIDA, 2012). Uma forma mais eficiente de apresentar esse conjunto de estatísticas é mapeá-las. Na Figura 16, o mapa de significância LISA exibe as regiões com estatística I local de Moran significativos para os setores censitários de casos confirmados de dengue para a cidade de São Carlos-SP.

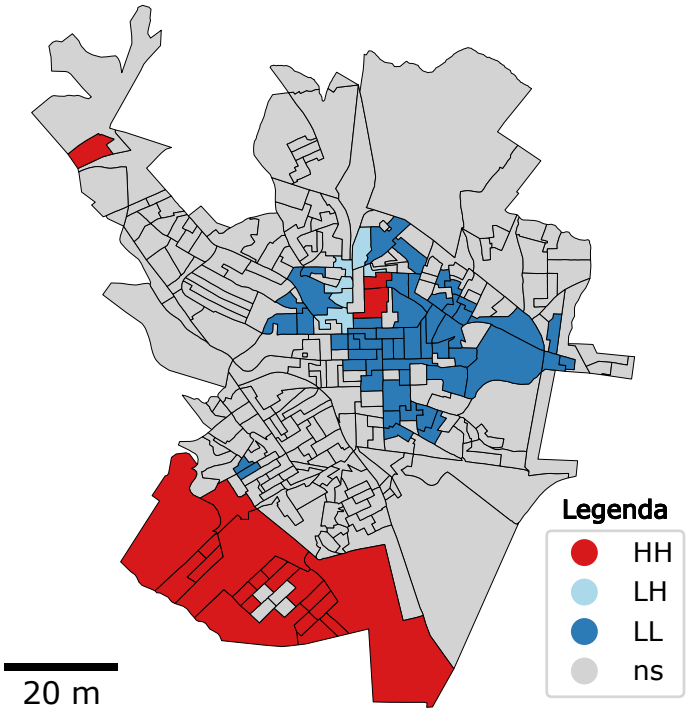
O mapa de *clusters* LISA combina a informação do diagrama de dispersão de Moran e a informação do mapa de significância das médias de associação local I_i. A Figura 17 apresenta os *clusters* que passaram no teste de significância estatística do I de Moran local, com os *clusters* em HH, HL, LH, LL e ns (não significante).

Figura 16 – Mapa de significância LISA para os setores censitários com casos confirmados de dengue da cidade de São Carlos-SP.



Fonte: Elaborada pelo autor.

Figura 17 – Mapa de *clusters* LISA para os setores censitários com casos confirmados de dengue da cidade de São Carlos-SP.



Fonte: Elaborada pelo autor.

2.3.4 Estatística de varredura espacial

A técnica estatística de varredura espacial foi desenvolvida por (KULLDORFF; NAGARWALLA, 1995), para identificar e localizar aglomerados de riscos presentes em uma determinada região de estudo. Kulldorff e Nagarwalla (1995) desenvolveu sua técnica, inspirado nos trabalhos de Openshaw *et al.* (1987) e Turnbull *et al.* (1990), reunindo as vantagens de cada técnica. Para cada local especificado, uma série de círculos de raios variados é construída. Cada círculo absorve os locais vizinhos mais próximos que estão dentro dele e o raio de cada círculo é definido para aumentar continuamente de zero até que uma porcentagem fixa da população total seja incluída. Para cada círculo, a hipótese alternativa é que exista um risco elevado de doença dentro do círculo em comparação com o exterior (KULLDORFF, 1997; KULLDORFF; RAND; WILLIAMS, 2006). O teste estatístico T_{KN} da varredura espacial é calculado pela Equação 2.9, a seguir:

$$T_{KN} = sup \left(\frac{O(Z)}{p(Z)} \right)^{n(Z)} \left(\frac{O(Z^c)}{p(Z^c)} \right)^{n(Z^c)} I \left(\frac{O(Z)}{p(Z)} > \frac{O(Z^c)}{p(Z^c)} \right), \tag{2.9}$$

onde Z^c indica todos os círculos exceto para Z, $O(.)$ e $P(.)$ são os números de observações dos casos e o tamanho da população em cada área respectivamente e $T(.)$ é uma função indicador. A simulação de Monte Carlo é realizada para comparar T_{KN}, com a distribuição de valores gerados sob a hipótese nula (PFEIFFER *et al.*, 2008).

Kulldorff (1997) implementou a estatística de varredura espacial em um programa de detecção de *cluster* chamado SaTScan[2], que procura *clusters* em conjuntos de dados usando dois modelos probabilísticos diferentes, um modelo de Bernoulli no qual casos e controles são comparados como variáveis *Booleanas* e um modelo de Poisson em que o número de casos é comparado com os dados da população de base e o número esperado de casos em cada unidade é proporcional ao tamanho da população em risco. Os centros dos círculos são definidos pelos dados de caso e controle/população ou pela especificação de uma matriz de coordenadas de grade. Os *clusters* secundários são calculados com base no grau de sobreposição permitido nos círculos do *cluster* e incluem as opções sem sobreposição geográfica e sem centros de *cluster* em outros *clusters* (PFEIFFER *et al.*, 2008; LOPES *et al.*, 2023a).

2.4 GIScience

A Ciência da Informação Geográfica (do inglês, *Geographic Information Science (GIScience)*) é uma área de estudo que busca formalizar os princípios geográficos para explorar aplicações científicas e relacionadas a políticas de informações geográficas. O objetivo da GIScience é revelar e analisar as complexas relações que indivíduos, organizações e a sociedade têm com as tecnologias de informações geográficas, bem como responder a perguntas fundamentais sobre o uso do GIS como uma ferramenta de suporte à decisão (MARK, 2003; MALCZEWSKI; RINNER, 2015).

Segundo Mark (2003) os componentes da GIScience abrangem os seguintes campos:

- **Modelos Cognitivos do Espaço Geográfico:** Esse componente se concentra na compreensão de como as pessoas percebem e entendem o espaço geográfico. Isso inclui a investigação de como as pessoas usam mapas e outras representações geográficas para navegar e tomar decisões, bem como a exploração de como diferentes culturas e contextos sociais afetam a percepção do espaço.

- **Métodos Computacionais para Representação de Conceitos Geográficos:** Já ese componente se concentra na criação de modelos e representações computacionais de dados geográficos. Inclui-se, dessa forma, a investigação de como os dados geográficos podem ser armazenados, organizados e analisados ao se utilizar tecnologias de informação e comunicação, além da exploração de como diferentes tipos de dados geográficos podem ser integrados e combinados para criar novos *insights*.

- **Geografias da Sociedade da Informação:** Tal componente foca nas implicações sociais e éticas do uso de informações geográficas. Investiga como as informações geográficas são coletadas, usadas e compartilhadas em diferentes contextos e trata da exploração de como

[2] <http://www.satscan.org>

as informações geográficas podem ser usadas para resolver problemas sociais e ambientais complexos.

- **Estatísticas espaciais:**Salienta-se que a estatística espacial é uma importante área de pesquisa com fortes vínculos com a GIScience e ela fornece métodos estatísticos formais para lidar com a autocorrelação espacial, como medi-la ou controlar seus efeitos ao conduzir análises estatísticas com base em dados para unidades espaciais. A estatística espacial pode ser usada para caracterizar alguns aspectos da qualidade dos dados.

GIScience procura responder a questões fundamentais sobre o uso de SIG. Essas perguntas são frequentemente feitas com referência ao SIG como uma ferramenta de suporte à decisão. Embora o SIG seja convencionalmente visto como um conjunto de ferramentas para entrada, armazenamento e recuperação, manipulação e análise e saída de dados espaciais, o sistema também contém um conjunto de procedimentos para apoiar as atividades de tomada de decisão (MALCZEWSKI; RINNER, 2015). De fato, GIS pode ser definido "como um sistema de apoio à decisão envolvendo a integração de dados referenciados espacialmente em um ambiente de solução de problemas" (MARK, 2003). Nesse contexto, o SIG é considerado um banco de dados digital de propósito especial no qual um sistema de coordenadas geográficas comum é o principal meio de armazenar e analisar os dados para obter informações para a tomada de decisões. Em última análise, o objetivo do uso do SIG é fornecer suporte para a tomada de decisões (MALCZEWSKI; RINNER, 2015).

2.5 Análise Hierárquica de Processos (AHP)

A definição fundamental dos vários problemas de tomada de decisão é o de racionalidade. De acordo com esse princípio, indivíduos e organizações seguem um comportamento de escolha entre alternativas, baseado em critérios objetivos de julgamento, cujo fundamento será satisfazer um nível desejado pré-estabelecido (HARKER, 1989).

O método AHP é uma metodologia para estruturação, medição e síntese de problemas de decisão multicritério (FORMAN; GASS, 2001). Esse visa simular a maneira como as pessoas pensam e é baseado em três princípios: decomposição, julgamento comparativo e síntese de prioridades. O princípio da decomposição exige que um problema de decisão seja decomposto em uma hierarquia que capture os elementos essenciais do problema. Já o princípio do julgamento comparativo requer a avaliação de comparações pareadas dos elementos em um determinado nível da estrutura hierárquica, com relação ao seu pai no próximo nível superior. Por fim, a comparação pareada é o modo de medição básico empregado no procedimento AHP. Percebe-se, dessa maneira, que o princípio de síntese toma cada uma das prioridades de escala de razão derivadas nos vários níveis da hierarquia e constrói um conjunto composto de prioridades para os elementos no nível mais baixo da hierarquia (ou seja, alternativas) (SAATY, 1987; SAATY, 1988;

SAATY, 1990). Dados esses princípios, o procedimento AHP envolve três etapas principais: (i) desenvolver a hierarquia AHP, (ii) atribuir pesos de importância a cada elemento da estrutura hierárquica ao usar o método de comparação pareada, e (iii) construção de uma classificação geral de prioridade (MALCZEWSKI; RINNER, 2015; MU *et al.*, 2017).

Segundo (SAATY, 1987),

> O método AHP por ser uma teoria geral de medição, ele é usado para derivar escalas de razão de comparações pareadas discretas e contínuas. Essas comparações podem ser tiradas de medições reais ou de uma escala fundamental que reflita a força relativa de preferências e sentimentos. O AHP tem uma preocupação especial com o afastamento da consistência, da sua medição e da dependência dentro e entre os grupos de elementos da sua estrutura e é utilizado na tomada de decisão multicritério, planejamento e alocação de recursos e na resolução de conflitos (SAATY, 1987).

A funcionalidade do método AHP é enumerada em oito usos principais. Permite ao decisor: 1) desenhar um formulário que represente um problema complexo; 2) medir prioridades e escolher entre alternativas; 3) medir a consistência; 4) prever; 5) formular uma análise de custo/benefício; 6) projetar planejamento para frente/para trás; 7) analisar a resolução de conflitos; 8) desenvolver alocação de recursos a partir da análise de custo/benefício (SAATY, 1987; SAATY, 1988; SAATY, 1990).

Outro ponto a se destacar é que, no método AHP, os julgamentos de comparação pareada são fundamentais para determinar a importância relativa dos elementos em questão. Para facilitar esse processo, é comum utilizar uma escala de 1 a 9, conhecida como escala fundamental de Saaty. No entanto, é importante ressaltar que atribuir números nessa não é uma mera atribuição arbitrária. A intensidade relativa dos elementos sendo comparados em relação a uma propriedade específica torna-se crucial (SAATY, 2008). Na Tabela 2 é apresentada a escala fundamental de valores proposto por (SAATY, 1987) para comparação pareada do método AHP.

Dessa maneira, no método AHP as comparações pareadas desempenham um papel fundamental. Inicialmente, é necessário estabelecer prioridades entre os principais critérios, avaliando-os em pares para determinar sua importância relativa e criando, assim, uma matriz de comparação pareada. Essas comparações são expressas por meio de números retirados da escala fundamental (Tabela 2). O número de comparações necessárias para uma matriz de ordem n, onde n é o número de elementos sendo comparados, é dado por $\frac{n(n-1)}{2}$, pois é recíproco e os elementos na diagonal são iguais a um. Entretanto, existem condições em que é possível usar um número menor de comparações e ainda obter resultados precisos. Durante as comparações, questiona-se em que medida o elemento à esquerda da matriz é considerado mais importante do que o elemento no topo da matriz em relação à propriedade em questão (SAATY, 1987).

Chan, Kwok e Duffy (2004) resume os passos recomendados para aplicação do AHP:

Tabela 2 – Escala de valores AHP para comparação pareada.

Intensidade de importância	**Definição e Explicação**
1	Importância igual Os dois fatores contribuem igualmente para o objetivo
3	Importância moderada Um fator é ligeiramente mais importante que o outro
5	Importância essencial Um fator é claramente mais importante que o outro
7	Importância demonstrada Um fator é fortemente favorecido e sua maior relevância foi demonstrada na prática
9	Importância extrema A evidência que diferencia os fatores é da maior ordem possível
2,4,6,8	Valores intermediários entre julgamentos Possibilidade de compromissos adicionais

Fonte: Adaptada de Saaty (1987).

1. Definir o problema e o que se procura saber. Expor as suposições refletidas na definição do problema, identificar partes envolvidas, checar como essas definem o problema e suas formas de participação no método AHP.

2. Decompor o problema desestruturado em hierarquias sistemáticas, do topo (objetivo geral) até o último nível (fatores mais específicos, usualmente as alternativas). Caminhando do topo para a extremidade, a estrutura do método AHP contém objetivos, critérios (parâmetros de avaliação) e classificação de alternativas (medição da adequação da solução para o critério). Cada nó é dividido em níveis apropriados de detalhes. Quanto mais critérios, menos importante cada critério individual se torna, e a compensação é feita pela atribuição de pesos para cada critério. É importante certificar-se de que os níveis estejam consistentes internamente e completos, e que as relações entre os níveis estejam claras.

3. Construir uma matriz de comparação paritária entre os elementos do nível inferior e os do nível imediatamente acima. Em hierarquias simples, cada elemento de nível inferior afeta todos os elementos do nível superior. Em outras hierarquias, elementos de nível inferior afetam somente alguns elementos do nível superior, o que requer a construção de matrizes únicas.

4. Fazer os $\frac{n^2-2}{2}$ julgamentos a partir dos valores da escala fundamenta de Saaty (Tabela 2).

5. Calcular o Índice de Consistência (IC) a partir da Equação 2.10. Se não for satisfatório, refazer julgamentos.

$$\text{índice de consistência} = IC = \frac{\lambda_{max} - n}{n-1}, \tag{2.10}$$

onde λ_{max} é o autovalor máximo da matriz de julgamento e n a dimensão dessa matriz.

6. Calcular a Razão de Consistência, que considera o IC e o Índice Randômico (IR), o qual varia com o tamanho n da amostra (Tabela 3).

$$\text{Razão de Consistência} = \frac{IC}{IRparan}, \tag{2.11}$$

Tabela 3 – Índice Randômico (IR) do método AHP.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0,00	0,00	0,58	0,90	1,12	1,24	1,32	1,41	1,45	1,49	1,51	1,48	1,56	1,57	1,59

Fonte: Saaty (1988).

7. Analisar as matrizes para estabelecer as prioridades locais e globais, comparar as alternativas e selecionar a melhor opção.

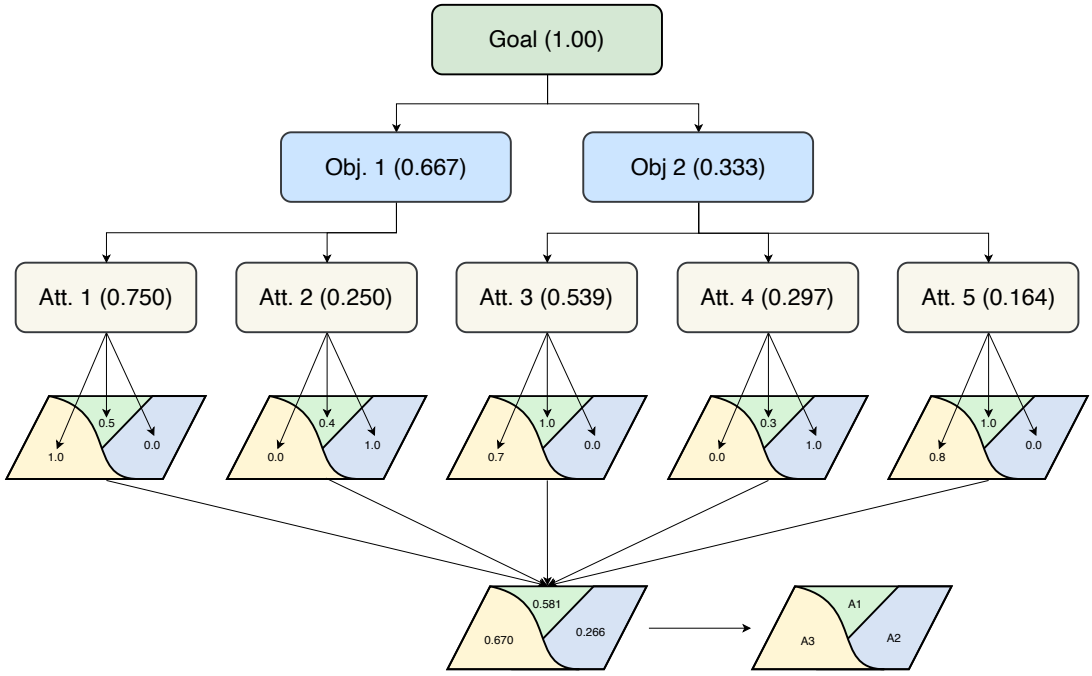
Devido a sua aplicabilidade, o método AHP é utilizado em uma ampla gama de situações envolvendo seleção entre alternativas concorrentes em um ambiente multiobjetivo, alocação de recursos escassos e previsão. Embora tenha essa característica a fundamentação axiomática do AHP delimita cuidadosamente o escopo do ambiente do problema (SAATY, 1988), pois baseia-se na estrutura matemática bem definida de matrizes consistentes e na capacidade de seu autovetor direito associado de gerar pesos verdadeiros ou aproximados (SAATY, 1987; SAATY, 1988; SAATY, 1990).

A literatura também aponta que existem diversos trabalhos na literatura para integrar GIS e AHP (MALCZEWSKI, 2006; MALCZEWSKI; RINNER, 2015). Segundo (MALCZEWSKI; RINNER, 2015) e a abordagem mais comun consiste em integrar o método AHP ao GIS como uma ferramenta para estimar os pesos associados às camadas do mapa de atributos/critérios. Uma vez que os pesos são estimados, esses são combinados com as camadas do mapa de atributos usando uma combinação linear ponderada (do inglês, *Weighted Linear Combination (WLC)*) (JANKOWSKI; RICHARD, 1994; EASTMAN; JIANG; TOLEDANO, 1998; MALCZEWSKI; RINNER, 2015).

A Figura 18 mostra um exemplo de aplicação do método AHP baseado em GIS. Um problema de avaliação de três parcelas de um determinado terreno ($A1$, $A2$ e $A3$) é primeiro decomposto em uma hierarquia que desce do objetivo geral ($Goal$) para os elementos mais específicos do problema: dois objetivos e cinco atributos. A importância relativa dos dois objetivos é avaliada com base na comparação pareada. Supondo que o Objetivo 1 seja duas vezes mais relevante que o Objetivo 2 e calculando os pesos dos objetivos da seguinte forma: $w_1 = (1/2)((1/1.5)+(2/3)) = 0,667$ e $w_2 = 0,333$ (ver Tabela 4a). Existem dois atributos associados ao Objetivo 1. A Tabela 4b mostra que o Atributo 1 é três vezes mais significativo que o Atributo 2 e consequentemente, $w_{1(1)} = 0,75$ e $w_{2(1)} = 0,25$. Os pesos dos atributos associados ao Objetivo 2 foram calculados de maneira semelhante (consulte a Tabela 4c). Como a razão de consistência

(CR) para cada uma das tabelas de comparação apreada é menor que < 0,10, os pesos podem ser usados para calcular o valor geral de cada alternativa usando através de uma WLC, da seguinte maneira: o valor global de $V(A1) = (0,6670,750,5) + (0,6670,250,4) + (0,3330,5391,0) + (0,3330,2970,3) + (0,3330,1641,0) = 0,581$. Os valores globais de $V(A2) = 0,266$ e $V(A3) = 0,670$ são calculados de maneira semelhante (MALCZEWSKI; RINNER, 2015).

Figura 18 – Exemplo de estrutura hierárquica de aplicação do método AHP baseado em GIS (Note que $A1$, $A2$ e $A3$ são alternativas, Obj. Objetivos, Att. Atributos e os mapas mostram valores dos atributos padronizados para cada mapa).



Fonte: Adaptada de Malczewski e Rinner (2015).

Uma das grandes vantagens de utilizar o método AHP para integrar com soluções baseadas em GIS é que ele fornece uma ferramenta para focar a atenção do tomador de decisão no desenvolvimento de uma estrutura formal para capturar todos os elementos consideráveis de uma situação de decisão (VARGAS, 1990; MALCZEWSKI; RINNER, 2015).

2.6 Meta-Heurísticas

Uma vez que os algoritmos computacionais enfrentam um desafio enorme quando se trata de fornecer soluções exatas para problemas de grande porte classificados como NP-Difíceis [3]. Independentemente de condições especiais ou propriedades particulares, a resolução de um problema NP-Difícil requer, em última instância, um consumo exponencial de tempo de processamento ou memória, a depender do tamanho dos dados de entrada. Problemas de grande

[3] Classe de problemas na qual as soluções não podem ser necessariamente verificadas em tempo polinomial (EIBEN; SMITH, 2015)

Tabela 4 – Comparações de pares de: (a) objetivos em relação à meta, (b) atributos em relação ao objetivo 1 e (c) atributos em relação ao objetivo 2.

(a)				
Alvo				
Objetivos	Obj. 1	Obj. 2	w_l	
Obj. 1	1	2	0.667	
Obj. 1	0.5	2	0.333	
Soma	1.5	3.0	1.000	
$CR = 0.00$				
(b)				
Objetivo 1	Att. 1	Att. 2	$w_{k(1)}$	
Att. 1	1	3	0.750	
Att. 2	0.33	1	0.250	
Soma	1.33	4.00	1.000	
$CR = 0.00$				
(c)				
Objetivo 2				
Atributos	Att. 3	Att. 4	Att. 5	$w_{k(2)}$
Att. 3	1	2	3	0.539
Att. 4	0.5	1	2	0.297
Att. 5	0.33	0.5	1	0.164
Soma	1.83	3.50	6.00	1.000
$CR = 0.01$				

Fonte: Adaptada de Malczewski e Rinner (2015).

porte, com um grande número significativo de variáveis, geralmente exigem um tempo de processamento ou uma quantidade de memória computacional proibitivamente alta, ou, até ambos, para obter uma solução exata. Diante da persistência desse desafio, cada vez mais se acredita que os computadores e métodos algorítmicos atuais apresentam uma limitação tecnológica inerente para alcançar soluções exatas eficientes, em geral, para problemas NP-Difíceis de grande porte e ainda mais complexos (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; GOLDBARG; GOLDBARG; LUNA, 2017).

Nos últimos anos, devido à inadequação das abordagens exatas, tem havido um crescente investimento e esforço no desenvolvimento e aprimoramento de estratégias aproximadas e eficientes para solucionar problemas NP-Difíceis de grande porte, tais procedimentos aproximativos são denominados de heurísticas [4] (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; GOLDBARG; GOLDBARG; LUNA, 2017). (GOLDBARG; GOLDBARG; LUNA, 2017) define heurística como "*uma técnica computacional aproximativa que visa alcançar uma solução avaliada como aceitável para um dado problema que pode ser representado em um computador que utilize um esforço computacional considerado razoável, sendo capaz de garantir,*

[4] Segundo (GIGERENZER; GAISSMAIER, 2011) uma heurística é uma estratégia que ignoram informações para tomar decisões de forma mais rápida e eficiente do que métodos mais complexos.

em determinadas condições, a viabilidade ou a otimalidade da solução." Em muitos casos, é esperado que as heurísticas se aproximem ou, até mesmo, alcancem os valores ótimos para a solução de problemas NP-Difíceis ou mais complexos, especialmente, quando são iniciadas a partir de uma solução viável próxima ao ótimo. No entanto, ao renunciar à garantia de encontrar a solução ótima do problema, os métodos heurísticos devem compensar oferecendo, no mínimo, eficiência computacional (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; GOLDBARG; GOLDBARG; LUNA, 2017).

Nos últimos anos, uma classe de estratégias abrangentes guia o processo de construção de heurísticas. Esses métodos são conhecidos como *heurísticas modernas* ou, mais amplamente difundido, *meta-heurísticas*. As meta-heurísticas representam uma arquitetura geral de regras que, com base em um tema comum, fornecem uma base sólida para o desenvolvimento de diversas heurísticas computacionais o que abrange uma ampla gama de aplicações e problemas. Essa abordagem flexível e adaptável tem sido amplamente explorada e comprovada como eficaz na busca de soluções de alta qualidade em problemas complexos e de difícil otimização (GOLDBARG; GOLDBARG; LUNA, 2017).

As regras que compõem o contexto meta-heurístico podem ser desenvolvidas de forma arbitrária, buscando inspiração em analogias com fenômenos físicos, químicos, biológicos, sociais ou, até mesmo, propriedades matemáticas. Essa flexibilidade permite a criação de abordagens inovadoras e criativas, adaptadas às características específicas dos problemas em questão. Além disso, essa diversidade de inspirações contribui para a robustez e adaptabilidade das meta-heurísticas, ao permitir que elas sejam aplicadas em uma ampla variedade de domínios e desafios (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; GOLDBARG; GOLDBARG; LUNA, 2017).

Os Algoritmos Genéticos (AG) são amplamente reconhecidos como uma das meta-heurísticas mais eficazes para resolver problemas de otimização (MITCHELL, 1998; GOLBERG, 1989; GOLDBERG, 2002). Proposto por Holland (1970) e inspirados pela teoria da evolução de Darwin (2004), os AGs utilizam conceitos como seleção natural, reprodução, *crossover* e mutação para explorar o espaço de buscas das soluções viáveis que visa encontrar soluções ótimas ou aproximadas para problemas complexos (MITCHELL, 1998; GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; GOLDBARG; GOLDBARG; LUNA, 2017).

Essa combinação de elementos da seleção natural com técnicas heurísticas permite que os algoritmos genéticos explorem eficientemente o espaço de busca de soluções em problemas NP-Difíceis de grande porte. Eles têm a capacidade de encontrar soluções aproximadas de alta qualidade, mesmo quando a busca pela solução ótima é inviável devido à complexidade do problema (MITCHELL, 1998).

Os AGs adotam um conjunto de terminologias para descrever situações e entidades relacionadas ao seu funcionamento. A coleção de soluções que são processadas é comumente chamada de **população**. Cada iteração do processo de otimização é considerada uma geração

e a sequência de várias populações reflete a **evolução** ao longo do tempo. Os elementos que compõem uma população são denominados **indivíduos** ou **cromossomos**. Os genes são os componentes básicos dos cromossomos e podem ser considerados as características elementares de uma solução. Eles estão localizados em posições específicas e cada um pode assumir um valor de um conjunto de possibilidades, chamadas **alelos** (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; GOLDBARG; GOLDBARG; LUNA, 2017).

A estrutura genérica de um AG pode ser observado no Algoritmo 1. Nele é possível visualizar o processo iterativo que gera sucessivas populações (passos 6 a 9).

Algoritmo 1 – Estrutura Genérica de um AG.

1: $t \leftarrow 0$
2: Gerar a população inicial $P(t)$
3: Avaliar os indivíduos de $P(t)$
4: **enquanto** critério de parada não for atingido **faça**
5: $t \leftarrow t+1$
6: Selecionar pais $P'(t)$ a partir de $P(t-1)$
7: Aplicar operadores genéticos a $P'(t)$ obtendo a nova população $P(t)$
8: Avaliar os indivíduos de $P(t)$
9: **fim enquanto**

À medida que o número de gerações aumenta, um AG converge gradualmente para regiões do espaço de procura onde se encontram soluções promissoras. A otimização termina quando um determinado critério de parada é atingido. Segundo Gaspar-Cunha, Takahashi e Antunes (2012), os critérios mais comuns são:

- Número limite de gerações atingido;
- Descoberta de uma solução com qualidade pretendida;
- Inexistência de melhoria durante um determinado período de tempo.

Ao atingir um critério de parada, o AG devolve o resultado da otimização. Existem duas alternativas para apresentação de resultados: devolver a melhor solução encontrada, ao longo da otimização, ou devolver um conjunto de indivíduos de qualidade elevada (por exemplo, devolver todos os indivíduos que fazem parte da última geração) (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; GOLDBARG; GOLDBARG; LUNA, 2017).

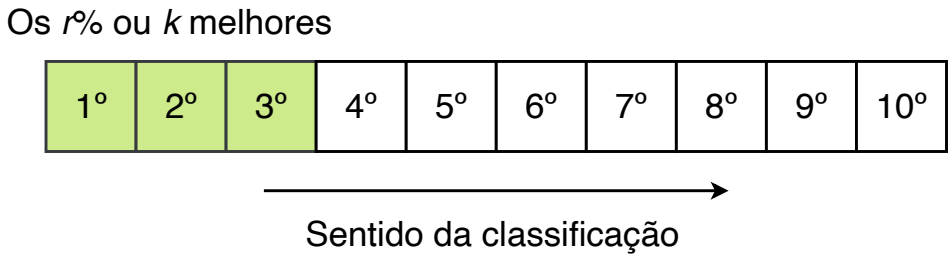
Seleção

Devido ao processo de seleção, a qualidade média dos elementos que compõem a população tende a aumentar ao longo do tempo. O operador de seleção emprega o princípio de sobrevivência do mais apto para escolher indivíduos de "alta qualidade" da população atual. A cada geração sucessiva, as soluções existentes são selecionadas com base em sua adequação;

quanto maior a qualidade do indivíduo, maior a probabilidade de ele ser selecionado para ser usado nos operadores genéticos (MALCZEWSKI, 2006; GOLDBARG; GOLDBARG; LUNA, 2017). De acordo com Talbi (2009), os procedimentos de seleção mais utilizados são: (i) elitismo, (ii) seleção por roleta e (iii) seleção de torneio.

No elitismo, os cromossomos são ordenados e escolhidos de acordo com sua classificação, os $r\%$ do comprimento da classificação ou os k melhores em cada geração. O objetivo desse é manter sempre o melhor indivíduo, da geração atual, na geração seguinte (MITCHELL, 1998). A Figura 19 exemplifica o critério exibindo a lista de classificação associada aos cromossomos.

Figura 19 – Classificação elitista.



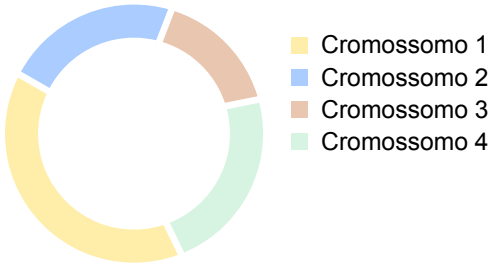
Fonte: Adaptada de Goldbarg, Goldbarg e Luna (2017).

Já o método de seleção por roleta é amplamente utilizado devido à sua facilidade de implementação. Nesse , cada cromossomo é distribuído em áreas proporcionais à sua aptidão. Portanto, os indivíduos com maior aptidão possuem uma maior probabilidade de serem selecionados. Uma abstração do método da roleta é apresentado na Figura 20, em que se deve executar os seguintes passos para selecionar um indivíduo:

1. Uma variável X recebe a somatória da aptidão de todos os indivíduos da população;
2. Uma variável aleatória r recebe um valor no intervalo $[0, X]$;
3. Cada indivíduo é percorrido, acumula-se o valor da aptidão de cada indivíduo em uma variável S;
4. Se $S \leq r$, então, o indivíduo corrente é selecionado, caso contrário continua a percorrer o próximo indivíduo.

Por fim, a seleção por torneio escolhe aleatoriamente m indivíduos da população para competirem pela chance de reproduzir ou fazer parte de uma população intermediária. O indivíduo i com melhor qualidade entre os m será selecionado. No caso de formação da população intermediária, esse processo é repetido até que sejam selecionados, por exemplo, o mesmo número de indivíduos da população original (MITCHELL, 1998). O Algoritmo 2 ilustra esse método, executando um torneio em que $m = 2$.

Figura 20 – Método da roleta.



Fonte: Adaptada de Goldbarg, Goldbarg e Luna (2017).

Algoritmo 2 – Seleção por torneio.

1: $n \leftarrow 0$
2: **repita** ▷ executa até que uma popula ção intermediária seja preenchida
3: Seleciona 2 indivíduos aleatoriamente
4: **se** $i_1.fitness > i_2.fitness$ **então** ▷ competição entre os indivíduos
5: i_1 é escolhido para população intermediária
6: **senão**
7: i_2 é escolhido para população intermediária
8: **fim se**
9: **até** $n =$ tamanho da população original

Outro exemplo de mecanismo de seleção é o método de seleção baseada em *rank* introduzido por Mitchell (1998). Essa estratégia utiliza as posições dos indivíduos quando ordenados de acordo com sua aptidão para determinar a probabilidade de seleção. Podem ser usados mapeamentos lineares ou não lineares para determinar a probabilidade de seleção. Uma forma de implementação desse mecanismo é simplesmente passar os n melhores indivíduos para a próxima geração (ZUBEN, 2000).

Operadores genéticos

Os operadores genéticos desempenham um papel fundamental na obtenção de novas soluções, ao mesmo tempo, em que procuram manter um nível adequado de diversidade no processo, ao equilibrar dois objetivos aparentemente conflitantes: o aproveitamento das melhores soluções e a exploração do espaço de busca (ZBIGNIEW, 1996; ZUBEN, 2000). Um AG inclui dois tipos principais de operadores genéticos: recombinação, também conhecida como *crossover*, e mutação. A implementação específica de cada um desses operadores depende da representação adotada para as soluções. Assim como a representação e a aptidão, esse é outro aspecto crucial que deve ser abordado com cuidado para maximizar a eficácia do processo de otimização (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; GOLDBARG; GOLDBARG; LUNA, 2017).

O operador de recombinação é o mecanismo de obtenção de novos indivíduos pela troca ou combinação dos alelos de dois, ou mais indivíduos. É o principal operador de reprodução dos

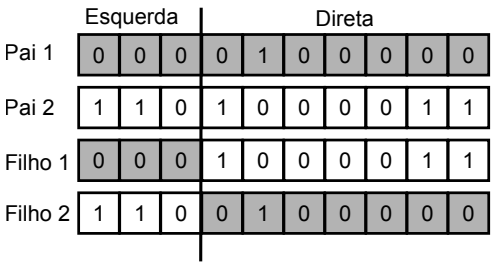
AGs (GOLBERG, 1989). Fragmentos das características de um indivíduo são trocadas por um fragmento equivalente oriundo de outro indivíduo. O resultado dessa operação é um indivíduo que combina características potencialmente melhores dos pais (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; GOLDBARG; GOLDBARG; LUNA, 2017). Dentre os possíveis operadores de cruzamento, pode-se descrever:

- **1-ponto:** Um tipo bastante comum de recombinação. Nessa operação, seleciona-se aleatoriamente um ponto de corte nos cromossomos, dividindo esse em uma partição à direita e outra à esquerda do corte. Cada descendente é composto pela junção da partição à esquerda (direita) de um pai com a partição à direita (esquerda) do outro pai (Figura 21).

- ***n*-pontos:** Outra recombinação, de n-pontos, divide os cromossomos em n partições, as quais são recombinadas. A Figura 22 ilustra um exemplo com 2 pontos de corte. Neste caso, o Filho 1 (Filho 2) recebe a partição central do Pai 2 (Pai 1) e as partições à esquerda e à direita dos cortes do Pai 1 (Pai 2).

- **Uniforme:** Outro tipo de recombinação muito comum é a recombinação uniforme, que considera cada gene independentemente, ao escolher de qual pai o gene do filho será herdado. Em geral, cria-se uma lista de variáveis aleatórias de distribuição uniforme em $[0, 1]$. Para cada posição, se o valor da variável aleatória for inferior a um dado P (usualmente 0, 5), o gene será oriundo do Pai 1; caso contrário, virá do Pai 2. O segundo filho é gerado pelo mapeamento inverso (ver Figura 23).

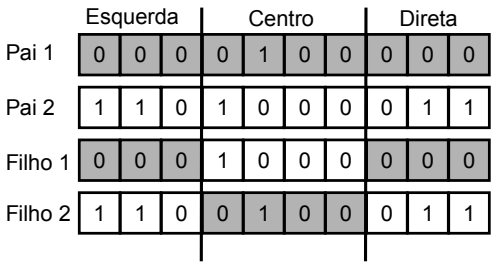
Figura 21 – Formas de recombinação: 1-ponto.



Fonte: Adaptada de Eiben e Smith (2015).

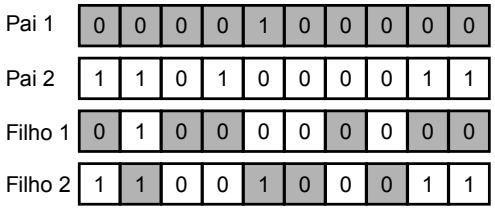
O operador de mutação é responsável por introduzir aleatoriedade e explorar novas regiões no espaço de busca. Ele modifica um ou mais genes de um cromossomo com base em uma **taxa de mutação** definida. A mutação desempenha um papel crucial na diversificação genética da população, permitindo que novas soluções sejam descobertas e evitando que o algoritmo fique preso em ótimos locais. No entanto, é comum atribuir valores baixos à taxa de mutação, geralmente na faixa de 0,001 a 0,1, para evitar mudanças drásticas que possam levar a uma deterioração significativa do desempenho (DEB, 2011).

Figura 22 – Formas de recombinação: 2-pontos.



Fonte: Adaptada de Eiben e Smith (2015).

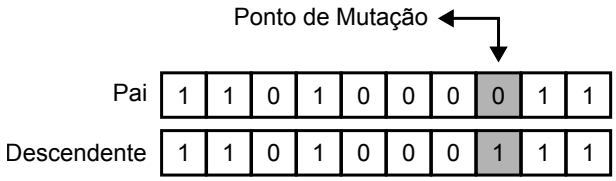
Figura 23 – Formas de recombinação: uniforme.



Fonte: Adaptada de Eiben e Smith (2015).

Considerando a codificação binária, o operador de mutação padrão simplesmente troca o valor de um gene em um cromossomo (GOLBERG, 1989). Assim, se o alelo de um gene selecionado é 1, o seu valor passará a ser 0 após a aplicação da mutação (Figura 24)

Figura 24 – Representação gráfica do operador de mutação.



Fonte: Adaptada de Eiben e Smith (2015).

No caso de problemas com codificação em ponto flutuante, os operadores de mutação mais populares são a mutação uniforme e a mutação Gaussiana (BÄCK; FOGEL; MICHALEWICZ, 2018; DEB, 2011). O operador para mutação uniforme seleciona aleatoriamente um componente $k \in 1,2,\ldots,n$ do cromossomo $x = [x_1,\ldots,x_k,\ldots,x_n]$ e gera um indivíduo $x' = [x_1,\ldots,x'_k,\ldots,x_n]$, em que x'_k é um número aleatório (com distribuição de probabilidade uniforme) amostrado no intervalo $[L_I, L_S]$, sendo L_I e L_S, respectivamente, os limites inferior e superior para o valor do alelo x_k. No caso da mutação Gaussiana, todos os componentes de um cromossomo x são modificados pela seguinte expressão:

$$x' = x + N(0,\sigma), \tag{2.12}$$

onde $N(0,\sigma)$ é um vetor de variáveis aleatórias Gaussianas independentes, com média zero e desvio padrão σ.

Representação real

A representação binária é a mais simples de reproduzir um cromossomo em algoritmos genéticos e frequentemente utilizada em diversos trabalhos (MALCZEWSKI; RINNER, 2015; GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; GOLDBARG; GOLDBARG; LUNA, 2017). Entretanto, diversos problemas do mundo real são abstraídos em valores contínuos, tornando, em muitos casos, inviável o uso da representação binária (DEB *et al.*, 2000). Por exemplo, a codificação de uma representação real em binária pode exigir o uso de uma longa sequência de valores binários e, consequentemente, tornar o processo evolutivo mais lento.

A representação real é uma abordagem alternativa que permite que os genes sejam representados por números reais, distribuídos em um intervalo contínuo de valores. Formalmente, essa representação pode ser expressa por meio de cromossomos pais, $p_1 = (p_1^1, p_1^2, \ldots, p_1^n)$ e $p_2 = (p_2^1, p_2^2, \ldots, p_2^n)$, e pelo cromossomo filho, $f = (f_1, f_2, \ldots, f_n)$. Nessa abordagem, p_1, p_2 e f são números reais ($\in \Re$) (SHISHIDO, 2018).

Segundo Shishido (2018), cromossomos codificados com valores reais podem usufruir de diversos, como:

- **Crossover média:** considerando cromossomos pais p_i^1 e p_i^2, o cromossomo descendente f é determinado por $f = (p_1 + p_2)/2$;
- **Crossover média geométrica:** considerando cromossomos pais p_i^1 e p_i^2, o cromossomo descendente f_i é resultante de $f_i = \sqrt{p_i^1 + p_i^2}$, onde $i = 1, 2, \ldots, n$;
- **Crossover BLX-α:** desenvolvido por Eshelman e Schaffer (1993), o *blend crossover*(cossover BLX-α) gera a partir de cromossomos pais p_i^1 e p_i^2 descendentes $f_i = p_i^1 + \alpha(p_i^2 - p_i^1)$, onde α é um número real entre 0 e 1;
- **Crossover linear:** a partir de dois pais, p_1 e p_2, esse operador gera três descendentes, f_1, f_2 e f_3, da seguinte forma: $f_1 = +0,5 \times p_1 + 0,5 \times p_2$, $f_2 = +1,5 \times p_1 - 0,5 \times p_2$ e $f_3 = -0,5 \times p_1 + 1,5 \times p_2$. Sendo que o descendente com melhor aptidão será escolhido.

A representação real também possibilita o desenvolvimento de diversos operadores de mutação, como:

- **Mutação uniforme:** substitui aletoriamente um gene desde que o valor esteja dentro do domínio do problema;
- **Mutação gaussiana:** é a alteração de um gene escolhido por um valor aleatório dentro de uma distribuição normal;

- **Mutação *creep*:** adiciona ao gene mutado um valor aleatório seguindo uma distribuição uniforme ou normal. Uma variação desse operador seria multiplicar o gene a ser mutado por um valor próximo de 1, produzindo uma peque perturbação;

- **Mutação por limite:** operador substitui o gene mutador por um dos limites do intervalo do problema;

- **Mutação não-uniforme:** substitui um gene por um valor extraído de uma distribuição não uniforme e,

- **Mutação não-uniforme múltipla:** todos os genes sofrem mutação seguindo as regas do operador de mutação não-uniforme.

Algoritmos steady-state

Nos AGs do tipo *steady-state*, não há mais a noção de gerações claramente definidas, na qual os indivíduos são totalmente substituídos a cada iteração. Em vez disso, tanto os pais quanto os descendentes coexistem e competem por um lugar na população. (JONG, 1975; WHITLEY *et al.*, 1989; WU; CHOW, 1995; GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; EIBEN; SMITH, 2015).

O Algoritmo 3 ilustra o modo de funcionamento de um AG do tipo *steady-state* no qual apenas um indivíduo pode ser substituído em cada iteração.

Algoritmo 3 – Algoritmo Genético *Steady-State*.

1: $t \leftarrow 0$
2: Gerar a população inicial $P(t)$
3: Avaliar os indivíduos de $P(t)$
4: **enquanto** critério de parada não for atingido **faça**
5: $t \leftarrow t+1$
6: Selecionar dois pais P_1 e P_2 a partir de $P(t-1)$
7: Criar um descendente D através da aplicação de operadores genéticos a partir de $P'(t)$
8: Avaliar o descendente D
9: Selecionais uma solução Z em $P(t-1)$ para substituição
10: Decidir se o descendente D substitui a solução Z em $P(t-1)$
11: **fim enquanto**

Nos AGs do tipo *steady-state*, os operadores genéticos geraram apenas um descendente, embora sejam criadas habitualmente duas soluções. Nesse passo pode-se assumir que a solução escolhida seja a de melhor qualidade ou que a escolha seja feita de forma aleatória (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012).

Já a escolha da solução Z, independente de domínio, pode ser feita utilizando vários critérios como substituição por qualidade, por idade, ou seja, a solução que está há mais tempo

na população, por semelhança ou mesmo aleatoriamente. É possível ainda considerar uma combinação de critérios para determinar a solução Z (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012).

No último passo do algoritmo *steady-state*, a decisão sobre a entrada do descendente D na população está normalmente relacionada com a sua qualidade, isto é, a solução é aceita se tiver melhor qualidade do que o indivíduo Z escolhido para ser substituído. Uma alternativa a esta abordagem é efetuar a substituição incondicional (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012).

2.6.1 Otimização Multiobjetivo

Métodos de otimização multiobjetivo ou análise de decisão multiobjetivo são uma classe de problemas no quais a qualidade de uma solução é definida pelo seu desempenho em relação a vários objetivos simultâneos, possivelmente conflitantes (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; EIBEN; SMITH, 2015; MALCZEWSKI; RINNER, 2015).

O objetivo em problemas multiobjetivos é encontrar um conjunto de soluções não dominadas, conhecido como conjunto de Pareto. Esse conceito foi introduzido por Pareto (1964) e é amplamente utilizado na área. Segundo Coello (2007) "um conjunto de soluções é dito ser de Pareto se não existir nenhuma solução desse conjunto que seja melhor em todos os objetivos do que outra solução desse conjunto". O conjunto de Pareto representa um conjunto de soluções que são consideradas ótimas, pois não é possível melhorar em um objetivo sem piorar em outro.

A solução para um problema multiobjetivo é chamada de solução Pareto-ótima ou Fronteira de Pareto. Diferentes soluções, Pareto-ótimas podem fornecer compromissos diferentes entre os objetivos. Por exemplo, algumas soluções podem ser melhores para um objetivo em particular, enquanto outras podem ser melhores para outros objetivos. O conjunto de Pareto completo é composto por todas as soluções Pareto-ótimas possíveis (COELLO, 2007; DEB, 2011).

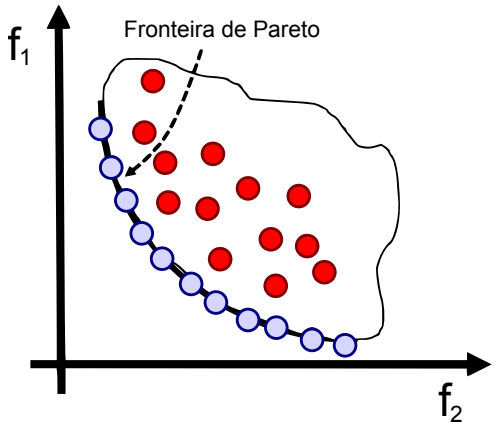
A ideia da Fronteira de Pareto é ilustrada na Figura 25 na qual cada ponto representa uma solução. Cada ponto é uma solução dominada por alguma não-dominada (ponto azul) da Fronteira de Pareto.

Existem na literatura várias abordagens para resolução de problemas multiobjetivo, através do uso de métodos matemáticos(HO; XU; DEY, 2010) e empregando computação evolutiva (DEB; DEB, 2013).

2.7 Considerações finais

Esse capítulo apresentou a fundamentação teórica necessária para compreensão da aplicação da tese. No próximo capítulo, será apresentada uma revisão da literatura, contendo

Figura 25 – Conjunto de soluções e a Fronteira de Pareto.



Fonte: Adaptada de Gaspar-Cunha, Takahashi e Antunes (2012).

um mapeamento sistemático para levantamento dos trabalhos relacionados à temática dessa produção.

CAPÍTULO

3

REVISÃO DA LITERATURA

3.1 Considerações iniciais

Este capítulo tem por objetivo apresentar um mapeamento sistemático utilizado como metodologia de pesquisa para busca e seleção do estado da arte sobre o uso de técnicas multicritério e algoritmos evolutivos em GIS.

3.2 Mapeamento sistemático

Nos últimos anos, houve um crescimento considerável no uso de métodos MCDM em soluções de problemas geoespaciais complexos. O campo de GIS-MCDA foi consistentemente adotado dentro da comunidade GIS. Os esforços para integrar (*multiple-criteria Decision Analysis (MCDA)*) ao GIS também foram reconhecidos como uma conquista significativa na expansão do MCDA para novas áreas de aplicação. Embora a principal motivação por trás dos esforços de pesquisa na integração do MCDA ao GIS venha da necessidade de expandir os recursos de suporte à decisão do GIS e tecnologias relacionadas, um significado igualmente importante é que as duas áreas distintas de pesquisa podem se beneficiar uma da outra. Por um lado, as técnicas e procedimentos GIS têm um papel importante a desempenhar na análise de problemas de decisão multicritério, pois oferecem recursos exclusivos para armazenar, gerenciar, analisar e visualizar dados geoespaciais para a tomada de decisões. Enquanto o GIS permite que analistas e tomadores de decisão pensem sobre as relações espaciais de uma maneira mais sofisticada e significativa do que seria possível de outra forma. Isso, por sua vez, permite desenvolver novas formas de pensar sobre alternativas de decisão e considerar novas soluções para problemas de decisão. Por outro lado, o MCDA pode melhorar a capacidade do GIS de lidar com problemas de decisão espacial de forma adequada. (MALCZEWSKI; RINNER, 2015).

A maneira usual para integrar os métodos MCDM e soluções envolvendo SIG (MALC-

ZEWSKI; RINNER, 2015) é a combinação linear ponderada (*weighted linear combination (WLC)*).

Chabuk *et al.* (2017) utilizam o método AHP para definir a importância hierárquica dos fatores físicos e ambientais que contribuem para a avaliação da potencial fragilidade ambiental da bacia hidrográfica do rio Jequitinhonha, Minas Gerais, Brasil. Hongoh *et al.* (2011) propõem outro trabalho com proposta semelhante, os autores utilizam uma abordagem baseada em MCDA para desenvolver modelos geoespaciais e ferramentas de apoio à decisão espacialmente explícitas para o gerenciamento de doenças transmitidas por vetores.

Vanolya, Jelokhani-Niaraki e Toomanian (2019) usam informações geradas por cidadãos como dados de linha de base para validar os resultados dos processos de MCDA e SIG. A validação proposta pelos autores é realizada por meio de índices espaciais específicos, o que inclui cobertura total, interseção geométrica, distâncias centrais e índices estatísticos. Rahman (2022) propõe investigar técnicas de visualização de dados para mapeamento de doenças para criar consciência sobre a doença para orientação de pacientes, profissionais de saúde e órgãos governamentais.

Yalew, Griensven e Zaag (2016) usam o método AHP para definir o grau de importância de múltiplos dados globais de diferentes fontes sobre agricultura por meio da plataforma Google em uma ferramenta chamada *framework* AgriSuit. Outro trabalho que utiliza o método AHP como técnica MCDA para auxiliar a tomada de decisão em dados espaciais é (DUAN *et al.*, 2022), em que os autores utilizam o método AHP para calcular o peso e gerar um mapa de resultados de perigos, exposição, vulnerabilidade e respostas emergenciais e capacidade de recuperação em áreas de risco de inundação urbana.

Visando levantar os trabalhos mais recentes sobre soluções GIS baseadas em algoritmos evolutivos multiobjetivos que utilizem múltiplas fontes de dados e heterogêneas, além de otimizarem a dependência espacial, foi realizado um Mapeamento Sistemático de Estudo (MSE) (*Systematic Mapping Study (SMS)*). MSE é um método científico capaz de identificar, interpretar e avaliar trabalhos de todas as pesquisas disponíveis relevantes para uma questão de pesquisa particular. Umas das razões para a realização de MSEs é que estes resumem as evidências existentes em relação a um tratamento ou tecnologia (KITCHENHAM *et al.*, 2009).

O MSE apresentado, nessa seção, segue as diretrizes de Kitchenham e Charters (2007) e Petersen *et al.* (2008). Petersen, Vakkalanka e Kuzniarz (2015) considera três fases principais: **(i) Planejamento -** engloba atividades de pré-revisão e estabelece um protocolo de revisão que define questões de pesquisa, critérios de inclusão e exclusão, fontes de estudo, *string* de pesquisa e mapeamento de procedimentos; **(ii) Condução -** procura e seleciona os estudos para extrair e sintetizar os dados deles; **(iii) Sumarização -** fase final para consolidar resultados e distribuí-los aos potenciais interessados. A motivação para tal método é fornecer uma possível síntese das evidências existentes, em relação ao tratamento ou à tecnologia, identificar tópicos para futuros pesquisadores e/ou para o desenvolvimento de base teórica relacionada a novas áreas de pesquisa

(KITCHENHAM *et al.*, 2009; PETERSEN; VAKKALANKA; KUZNIARZ, 2015).

3.2.1 Planejamento

Nesta fase, o protocolo de revisão contém: (i) questões de pesquisa; (ii) estratégia de busca;(iii) critérios de inclusão e exclusão e (iv) processo de extração dos dados e metodologia para a síntese dos dados foram definidos.

Questão de pesquisa

O principal objetivo deste MSE é identificar os trabalhos mais recentes sobre soluções GIS baseadas em algoritmos evolutivos multiobjetivos que utilizem múltiplas fontes de dados e heterogêneas, além de otimizarem a dependência espacial.

Estratégia e processo de busca

Para a recuperação dos estudos foi realizado um processo de seleção, ao qual foram abordados os seguintes aspectos: (i) termos e definição de *string* de busca; (ii) seleção de fonte para pesquisa; (iii) definição de critérios de inclusão e exclusão; e (iv) forma de armazenamento de dados.

String de busca

O processo de construção da *string* de busca considerou as seguintes palavras-chave em inglês: "geographic information system", "geographic information science", geostatistics, "spatial statistics", "multicriteria decision analysis", "multicriteria decision-making", "multiobjective optimization", "multi-objective optimization", "evolutionary algorithm", "genetic algorithm".

A *string* de busca elaborada (Figura 26) levou em consideração três áreas: **Problemas Espaciais e GIS** e **Técnicas de Decisão e Análise Multicritério** e **Otimização multi-objetivo com Computação Evolucionária**.

Figura 26 – *String* de busca.

```
("geographic information system" OR "geographic information science" OR gis OR
geostatistics OR "spatial statistics") AND ("multicriteria decision analysis" OR
"multicriteria decision-making" OR mcda OR mcdm) AND ("multiobjective
optimization" OR "multi-objective optimization" OR "genetic algorithm" OR ga OR
"evolutionary algorithm" OR ea)
```

Fonte: Elaborada pelo autor.

Para selecionar as bases de dados foram considerados os critérios discutidos por Dieste e Padua (2007). As seguintes bases de dados foram utilizadas: ACM Digital Library, EI Compendex, IEEE Xplore, Science Direct, Scopus e Web of Science (Tabela 5).

Tabela 5 – Bases de Dados utilizadas.

Base de Dados	Endereço eletrônico
ACM Digital Library	<http://dl.acm.org/>
Compendex	<https://www.engineeringvillage.com/>
IEEE Xplore	<http://ieeexplore.ieee.org/>
Science Direct	<https://www.sciencedirect.com/>
Scopus	<http://www.scopus.com/>
Web of Science	<https://webofknowledge.com/>

Fonte: Elaborada pelo autor.

Critérios de inclusão e exclusão

Para selecionar os trabalhos a serem analisados foram utilizados: um critério de inclusão (CI) e sete critérios de exclusão (CE) para inserção desses no mapeamento sistemático, definidos a seguir: os seguintes critérios de inclusão e exclusão: (**CI-1:**) Soluções GIS baseadas em algoritmos evolutivos multiobjetivos. Os critérios de exclusão foram: (**CE-1:**) O artigo não foi publicado nos últimos cinco anos; (**CE-2:**) O estudo não está escrito em inglês; (**CE-3:**) No caso de estudos duplicados, o mais completo é considerado; (**CE-4:**) O estudo não relaciona técnicas ou abordagens de problemas de otimização multiobjetivo em problemas espaciais; (**CE-5:**) O estudo não detalha a implementação utilizada; (**CE-6:**) O estudo primário é uma tabela de conteúdos, descrição de um curso curto, tutorial, palestras, formulário de direitos autorais ou resumo de um evento (por exemplo, uma conferência ou um *workshop*); **CE-7:** O estudo está publicado na literatura cinza [1].

Processo de Busca e seleção dos estudos primários

O processo de busca dos estudos primários foi realizado aplicando a *string* de busca nas bases de dados, definidos na Tabela 5, fazendo uso dos próprios mecanismos de busca avançada disponíveis em cada uma delas.

Como resultado dessa etapa de busca, foram encontrados 291 artigos, entre os quais a Science Direct retornou um conjunto maior de estudos (166). ACM Digital Libary, EI Compendex, IEEE Xplore, Scopus e Web of Science retornaram 40, 25, 2, 21 e 37, respectivamente. A ferramenta *StArt* (HERNANDES *et al.*, 2012) foi utilizada para remoção dos trabalhos duplicados (42 artigos).

A seleção dos trabalhos relevantes para o mapeamento sistemático foi realizada em três fases, nas quais se aplica, sucessivamente, os critérios de inclusão e exclusão definidos anteriormente, em trechos distintos dos documentos retornados pela busca inicial. Desta forma,

[1] Literatura cinza utiliza materiais/pesquisa disponibilizados por organizações que não pertencem à publicação comercial acadêmica ou tradicional

por meio da execução de cada uma das etapas do processo, é possível eliminar os estudos não condizentes com o objetivo da revisão.

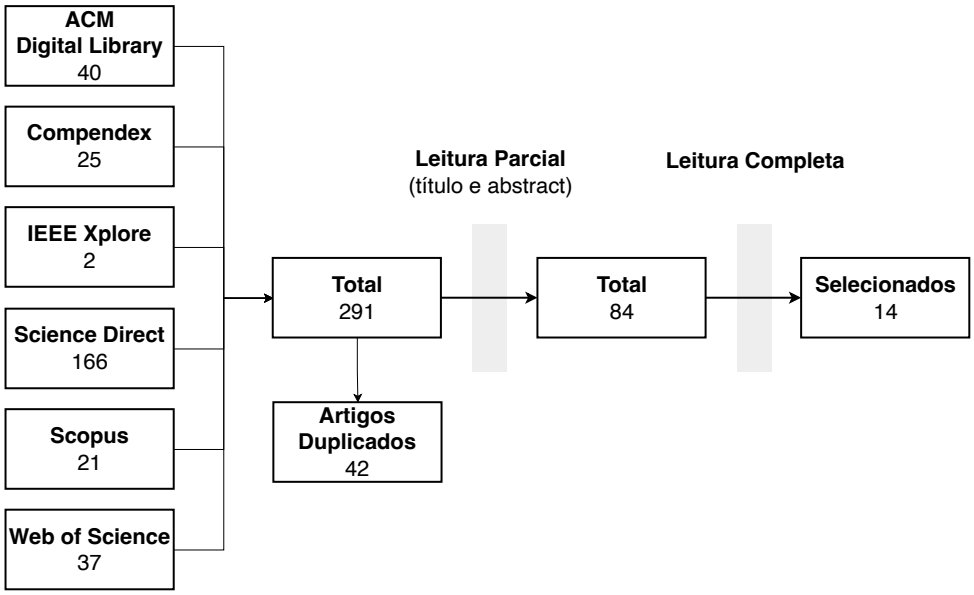
a. **Fase I: Leitura de títulos e resumos:** nessa etapa, foram aplicados os critérios de inclusão e exclusão em todos os trabalhos candidatos à análise contidos retornados pela busca inicial.

b. **Fase II: Leitura das seções de introdução e conclusão:** nessa etapa, foi realizada a leitura das seções de introdução e conclusão dos trabalhos selecionados na etapa anterior;

c. **Fase III: Leitura completa dos trabalhos:** nessa etapa, os estudos remanescentes são analisados completamente, sendo considerado todo seu conteúdo. Após a realização dessa etapa, obtém-se a relação final dos textos considerados na revisão.

A Figura 27 ilustra o processo de seleção dos estudos, por critérios de inclusão e exclusão, em cada uma das fases realizadas.

Figura 27 – Distribuição dos estudos primários (fase de condução).



Fonte: Elaborada pelo autor.

3.2.2 Extração e síntese de dados

Para a extração e síntese dos dados dos trabalhos selecionados, foi utilizado um formulário para responder a questão de pesquisa, além de extrair outras informações conforme ilustrado pela Tabela 6. A pesquisa foi realizada entre abril e junho de 2023 e os dados extraídos foram documentados e estão disponíveis em <https://bit.ly/sme-tese>.

Um mapeamento sistemático requer um esquema de classificação (PETERSEN *et al.*, 2008). Consideraram-se diferentes facetas, uma para cada questão de pesquisa. Examinaram-se

Tabela 6 – Conteúdo do formulário de extração dos dados para os trabalhos selecionados.

	Atributos
Metada	ID.
Conteúdo	Título, autor, ano, palavras-chave, resumo, doi tipo documento (i.e. conferência ou *jornal*), origem, base de dados.
Dados extraídos	1) método; 2) formato dos dados; 3) múltiplas fontes; 4) Dados heterogêneos. 5) Otimiza a dependência Espacial

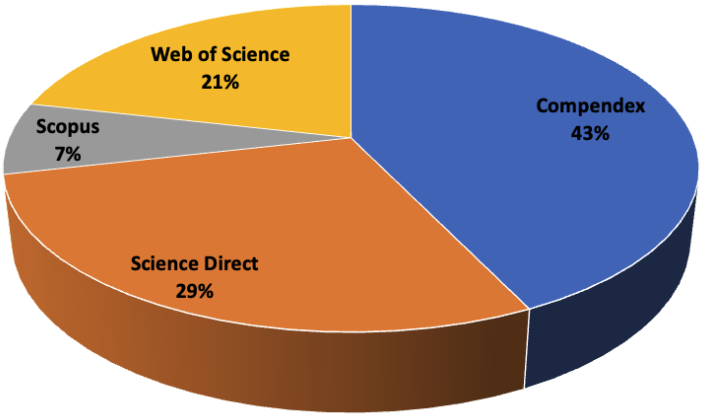
Fonte: Elaborada pelo autor.

apenas as principais descobertas encontradas como a utilização de métodos MCDA e multiobjetivo em GIS relatado pelos estudos selecionados. As categorias que compreendem outras facetas foram definidas segundo duas abordagens: (i) basear em categorias já consideradas na literatura; e (ii) levar os estudos selecionados em consideração. A seguir, as categorias dessas facetas são apresentadas na seção 3.2.3.

3.2.3 Resultados

Esta seção discute uma visão geral dos 14 estudos primários selecionados para o mapeamento sistemático. A Figura 28 apresenta a disposição dos estudos primários selecionados considerando cada base de dados utilizada. É possível verificar que foram selecionados em nosso estudo apenas estudos primários da IE Compendex com seis artigos (43%), Science Direct com quatro artigos (29%), Web of Science com três artigos (21%) e Scopus com apenas um trabalho (7%).

Figura 28 – Trabalhos selecionados por base de dados.

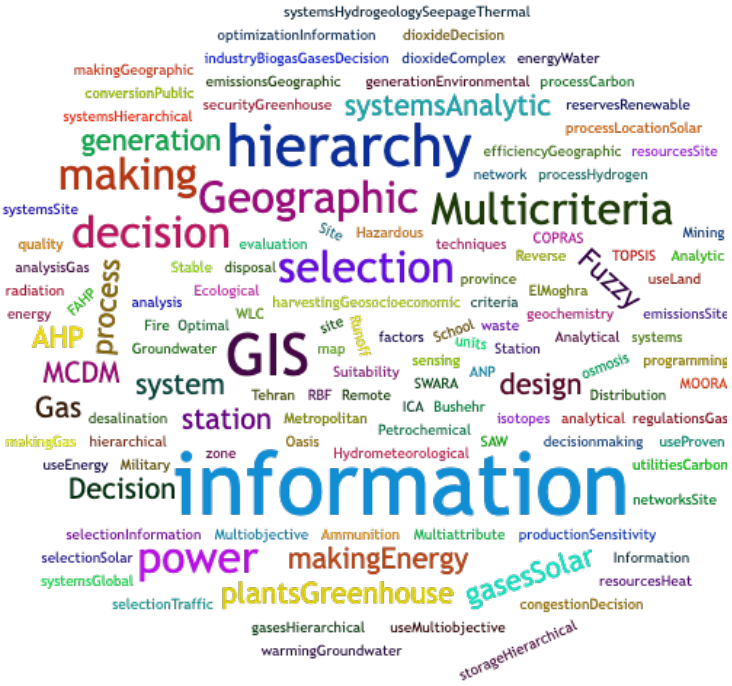


Fonte: Elaborada pelo autor.

Todos os trabalhos selecionados tiveram foram publicados em periódicos (*journal*)

No que se refere às palavras-chave extraídas, foi proposto um gráfico do tipo *word cloud* para ilustrar os atributos de indexação mais comuns. A Figura 29 representa a frequência em que cada uma das palavras-chave foi utilizada para caracterizar as publicações analisadas. Assim, o tamanho relativo de cada uma das expressões ilustradas representa o quão frequente ela foi utilizada.

Figura 29 – Relação das *keywords* mais frequentes.



Fonte: Elaborada pelo autor.

É possível observar que as palavras-chave mais comuns estão relacionadas à GIS, *decicion making* e multicritério.

3.2.4 Discussão das questões de pesquisa

A discussão sobre os estudos primários selecionados considera a questão de pesquisa proposta pelo MSE. A Tabela 6 representa o formulário usado para extrair as respostas para a questão de pesquisa. Os artigos selecionados por este mapeamento podem ser vistos na Tabela 7. Esta Tabela contém todos os artigos selecionados e suas referências bibliográficas e em <http://bit.ly/sm-psp-hpc> encontram-se os detalhes do esquema de classificação.

Tabela 7 – Lista final de estudos primários selecionados para extração de dados.

Título	Tipo do Dado	Método Pesos	Múltiplas Fontes	Dados Heterogêneos	Otimiza Dependência Espacial	Referência
GIS-based analytical analysis for selecting potential runoff harvestingsites: the case study of Amman-Zarqa Basin	raster	Weighted overlay analysis	sim	sim	não	(ODEH *et al.*, 2023)
GIS-Based SWARA and Its Ensemble by RBF and ICA Data-Mining Techniques for Determining Suitability of Existing Schools and Site Selection of New School Buildings	vetor	SWARA	não	não	não	(PANAHI *et al.*, 2019)
Compilation of a model for hazardous waste disposal site selection using GIS-based multi-purpose decision-making models	vetor	ANP	sim	sim	não	(DANESH *et al.*, 2019)
Optimal Hydrometeorological Station Network Design Using GIS Techniquesand Multicriteria Decision Analysis	raster	AHP	não	não	não	(FELONI; KARPOUZOS; BALTAS, 2018)
Potential aquifer mapping for cost-effective groundwater reverse osmosis desalination in arid regions using integration of hydrochemistry, environmental isotopes and GIS techniques	raster	AHP	sim	sim	não	(BOSELA *et al.*, 2022)
GIS-based fuzzy multi-criteria approach for optimal site selection of fire stations in Istanbul, Turkey	vetor	Fuzzy AHP	sim	sim	não	(NYIMBILI; ERDEN, 2020)

Continua na próxima página

Tabela 7 – Continuação da página anterior

Título	Tipo do Dado	Método Pesos	Múltiplas Fontes	Dados Heterogêneos	Otimiza Dependência Espacial	Referência
Solving an ammunition distribution network design problem using multi-objective mathematical modeling, combined AHP-TOPSIS, and GIS	raster	AHP-TOPSIS	sim	sim	não	(AKGÜN; ERDAL, 2019)
GIS based multi-criteria decision making for solar hydrogen production sites selection in Algeria	vetor	AHP	sim	sim	não	(MESSAOUDI *et al.*, 2019)
Territorial planning for photovoltaic power plants using an outranking approach and GIS	raster	AHP	sim	sim	não	(MARQUES-PEREZ *et al.*, 2020)
Complex power-to-gas plant site selection by multi-criteria decision-making and GIS	raster	AHP	sim	sim	não	(SOHA; HARTMANN, 2022)
Evaluation of GIS based Ranking and AHP methods in selecting the most suitable site: A case study in Kayseri, Turkey	raster	AHP	sim	sim	não	(GÜNEN, 2021)
A GIS-based MCDM approach for the evaluation of bike-share stations	raster	AHP	sim	sim	não	(KABAK *et al.*, 2018)

Continua na próxima página

Tabela 7 – Continuação da página anterior

Título	Tipo do Dado	Método Pesos	Múltiplas Fontes	Dados Heterogêneos	Otimiza Dependência Espacial	Referência
Indicator of suitability for evaluating the aquifer thermal energy storage using the GIS-based MCDA technique in the Halabja-Khurmal sub-basin	raster	AHP	sim	sim	não	(RAUF; ALI; AL-ANSARI, 2023)
Mining zone determination of natural sandy gravel using fuzzy AHP and SAW, MOORA and COPRAS methods	vetor	Fuzzy AHP	não	não	não	(AJRINA *et al.*, 2020)

Fonte: Dados da pesquisa.

O trabalho proposto por Odeh *et al.* (2023) tem o objetivo de selecionar os locais otimizados para captação de água de chuva através de SIG em uma região noroeste da Jordânia. Os autores utilizam um conjunto de 12 variáveis, chamado no trabalho de critérios, retiradas do (*United States Geological Survey (USGS)*). A definição do grau de relevância de cada variável utilizada no artigo é especificada por pesquisas anteriores e utilizada como fator de ponderação em uma combinação linear ponderada (*Weighted Overlay Analysis (WOA)*). O trabalho primeiro agrupa os critérios segundo quatro critérios (baixa, média, alta e muito alta) em termos de adequação da captação de água e aplica uma abordagem booleana para determinar se os critérios são aptos e/ou não quanto à captação de água. Após a aplicação da técnica booleana, é aplicada uma combinação linear ponderada aos critérios que foram considerados aptos, o que gera um mapa final resultante de adequação quanto à captação de água na área de estudo. Ao final, o mapa resultante é dividido em cinco categorias: não adequada, baixa adequação, média adequação, alta adequação e muito alta adequação.

Em (PANAHI *et al.*, 2019), os autores propõem a combinação do método de tomada de decisão (*Stepwise Weight Assessment Ratio Analysis (SWARA)*) com técnicas de mineração de dados *Radial Basic Function (RBF)* e *Imperial Competitive Algorithm (ICA)* para seleção de novos edifícios escolares, apoiados em SIG. O trabalho calcula, inicialmente, os pesos com o método SWARA e depois utiliza RBF e ICA para interpolar os fatores considerado. Após a definição dos pesos para cada classe e subclasse, é utilizada uma combinação linear ponderada, sendo os pesos o fator de ponderação.

O principal objetivo do trabalho proposto por Danesh *et al.* (2019) é encontrar o melhor local para descarte de resíduos perigosos em Bushehr, um dos maiores centros industriais (para poluentes químicos) do Irã, enfatizando critérios ambientais eficazes e utilizando o método *Analytical Network Process (ANP)*, proposto por Saaty *et al.* (1996) para melhorar o método AHP para definição do fator de importância de cada critério considerado. Após a definição dos pesos pelo método ANP, é utilizada uma combinação linear ponderada para gerar o mapa final com as zonas de melhor local para descarte de resíduos.

Em (FELONI; KARPOUZOS; BALTAS, 2018), os autores apresentam uma abordagem baseada em GIS e multicritério para a otimização de uma rede de estações hidrometereológicas , que visa estabelecer uma rede de estações ótima na região de Florina, no norte da Grécia. Os autores utilizam o método AHP para definição do grau de importância de cada critério utilizado e uma combinação linear ponderada pelos pesos gerados para compor o mapa final contendo a região com os melhores locais para instalação das estações hidrometereológicas .

O trabalho proposto por Bosela *et al.* (2022), visa encontrar locais econômicos para a construção de usinas de dessalinização para resolver a escassez de água doce em uma região árida no deserto ocidental do Egito. Para atingir esse objetivo os autores utilizam quatro critérios principais relacionados com o custo da dessalinização de águas subterrâneas: fparâmetros hidráulicos do aquífero, características hidrogeoquímicas, indicadores de recarga baseados em

traçadores isotópicos e outros parâmetros diversos relacionados à acessibilidade do local e questões ambientais. Esses quatro critérios principais incluem quatorze subcritérios; profundidade da água, espessura saturada do aquífero, transmissividade, sólidos totais dissolvidos, nível ferro, dureza total da água subterrânea, sulfato de cálcio, sílica, estrôncio, oxigênio-18, radiação solar global, declividade do terreno e distância das estradas. Os autores utilizaram o método AHP para organizar e definir os pesos dos critérios e subcritérios. Após definição dos pesos é criado um mapa de aptidão final é criado a partir da integração dos critérios por meio de uma combinação linear ponderada. Os melhores locais para a construção da usina de dessalinização são avaliados no mapa de aptidão final utilizando o método *Groundwater Aquifer Level Distance Impact Thickness (GALDIT)*, um método de classificação numérica baseado em técnicas de sobreposição e índice (LOBO-FERREIRA *et al.*, 2005; FERREIRA; CHACHADI, 2005; MOGHADDAM; JAFARI; JAVADI, 2017). O mapa de aptidão final é organizado em três níveis (baixa, média e alta suscetibilidade).

Nyimbili e Erden (2020) propõem o uso da extensão *fuzzy* do método AHP, integrado com uma solução utilizando SIG para otimizar o local de novas estações de combate a incêndios para uma região de Istambul. Esta abordagem *fuzzy* proposta simula os julgamentos subjetivos de especialistas para as preferências dos seis critérios avaliados para o mapeamento de adequação do quartel de bombeiros e, assim, contabiliza a incerteza de valores de comparação nítidos por meio de números *fuzzy* triangulares. Os pesos dos critérios avaliados a partir deste procedimento foram usados em uma análise de sobreposição ponderada das camadas do mapa de critérios reclassificados para gerar um mapa de adequação do quartel de bombeiros. Esses pesos resultantes do critério *fuzzy* AHP foram validados usando outra técnica MCDM, chamada *Best-Worst Method (BWM)* e considerados comparáveis e consistentes. Com base em uma avaliação minuciosa nas áreas representadas por valores de classe variando de 3 a 5 no mapa de adequação, um total de 34 novos locais de quartéis de bombeiros foram selecionados complementando os 121 quartéis de bombeiros existentes. Além disso, uma análise de priorização de baixo, médio a alto foi realizada para planejar as fases para a construção de novos quartéis de bombeiros em vista das necessidades orçamentárias concorrentes e restrições de recursos. A metodologia para alcançar isso foi proposta e modelada para aprimorar o processo de tomada de decisão em estudos de seleção de locais de quartéis de bombeiros urbanos.

Akgün e Erdal (2019) propõem uma metodologia que utiliza modelagem matemática multiobjetivo, o método AHP, a técnica (*Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)*) e SIG para resolver o problema de distribuição de munição de nível estratégico (*Ammunution Distribution Network Design Problem (ADNDP)*). Os autores utilizam um modelo matemático multiobjetivo para determinar as localizações e as atribuições de serviço dos depósitos considerando dois objetivos, custos de transporte e pontuações de risco dos principais depósitos. A pontuação de risco, de um local de depósito, indica o quão vulnerável a localização é a interrupções, sendo que essa pontuação é determinada por uma análise AHP-TOPSIS que combinada TOPSIS para calcular as pontuações de risco e AHP para medir os pesos

necessários para TOPSIS. A análise por meio do SIG é conduzida para determinar os locais de depósito potenciais usando camadas de mapa com base em critérios espaciais.

Em (MESSAOUDI *et al.*, 2019), é desenvolvido uma estrutura integrada para avaliar a adequação da terra para a produção de hidrogênio a partir da seleção do local de energia solar que combina MCDM com SIG. No SIG, são considerados dois tipos de critérios: restrições e critérios de ponderação. Os critérios de condicionantes permitirão reduzir a área de estudo, descartando as áreas que impedem a implementação da instalação de sistemas solares de produção de hidrogênio. Esses serão obtidos a partir da legislação (uso do solo, corpos d'água, hidrovias, rodovias, ferrovias, linhas elétricas e também o seu entorno). Os critérios de ponderação serão escolhidos conforme a demanda de hidrogênio, potencial de produção solar de hidrogênio, modelos digitais de elevação, declividade, proximidade de estradas, ferrovias e linhas de energia. Os autores utilizam o método AHP para ponderar cada critério, visando avaliar potenciais locais para a localização de um sistema de instalação de produção solar de hidrogênio. Após a definição da contribuição de cada critério, é realizada uma combinação linear ponderada para obtenção de um mapa final contendo a adequação dos locais para o sistema de instalação de produção de hidrogênio movido à energia solar. O mapa final é agrupado em quatro categorias como "adequação muito baixa", "adequação baixa", "adequação moderada" e "adequação alta" com um método de classificação de intervalo manual.

Em (MARQUES-PEREZ *et al.*, 2020), os autores propõem um estudo sobre o planejamento territorial para usinas de energia fotovoltaica usando uma abordagem de classificação e SIG. O objetivo do estudo é identificar locais ótimos para o desenvolvimento de usinas solares, o que é crucial para a viabilidade econômica desses projetos e o uso sustentável da terra. O estudo é aplicado à Comunidade Valenciana, uma região europeia situada no leste da Espanha. Os autores utilizam, no estudo, um total de 12 critérios e 15 subcritérios agrupados em três categorias: econômica, social e ambiental. Cada critério é avaliado com base em subcritérios específicos. Por exemplo, o critério de radiação solar é avaliado com base em subcritérios como a intensidade da radiação solar, a variação sazonal da radiação solar e a variação diária da radiação solar. Os autores utilizam *Preference Ranking Organization METHod for Enrichment of Evaluations (PROMETHEE)* e AHP para avaliar e classificar as áreas com base em critérios e subcritérios específicos. O estudo utiliza uma metodologia de avaliação multicritério usando PROMETHEE e AHP para analisarr e classificar as áreas com base nesses critérios e subcritérios. O método PROMETHEE é usado para avaliar e classificar as áreas com base em critérios qualitativos e quantitativos. Contudo, o método AHP é usado para avaliar e classificar as áreas com base em critérios hierárquicos e subcritérios. Após a definição do peso de cada critério e subcritério, é utilizada uma combinação linear ponderada para criação de um mapa que mostra uma classificação de áreas com alto potencial para o desenvolvimento de usinas solares. O mapa final gerado mostra as áreas com maior potencial para o desenvolvimento de usinas solares, o que é crucial para a viabilidade econômica desses projetos e o uso sustentável da terra.

Soha e Hartmann (2022) apresentam uma metodologia de adequação e seleção de locais baseada em SIG, considerando serviços de utilidade pública como redes de energia, gás e água, além de outros critérios comuns de adequação de locais em Borsod-Abaúj-Zemplen, Hungria. Esses fatores foram ponderados através do método AHP. Após a definição do peso de cada critério e subcritério, é utilizada uma combinação linear ponderada para definição de um mapa final.

Em (GÜNEN, 2021), os autores utilizam o método AHP integrado com SIG para seleção do local mais adequado para usinas de energia solar fotovoltaica em Kayseri, Turquia. Os autores utilizam três métodos de classificação e ponderação: *rank sum*, *rank reciprocal weights* e *rank order centroid weights*, além do método do AHP para definição da contribuição de cada critério utilizado no estudo. Após a definição da contribuição de cada critério, é utilizada uma combinação linear pondera para definição de um conjunto de mapas de adequação que mostram os locais mais adequados para a instalação de usinas de energia solar fotovoltaica em Kayseri, Turquia. Esses mapas são divididos em cinco níveis de adequação: excelente, bom, justo, baixo e ruim.

Kabak *et al.* (2018) e seus colaboradores apresentam uma abordagem utilizando SIG baseada em AHP e Análise Multicritério de Otimização e Ranking (*Multi-objective optimization by ratio analysis (MOORA)*) para avaliar a adequação de locais para estações de bicicletas compartilhadas. O primeiro passo do método proposto consiste em coletar de dados sobre a localização das estações de bicicletas compartilhadas existentes e as características do ambiente urbano em que estão localizadas. Esses dados são inseridos em um SIM para que seja feita uma análise espacial dos dados. Em seguida, os autores utilizam o método AHP para atribuir pesos às diferentes variáveis que afetam a adequação do local para uma estação de bicicletas compartilhadas. Essas variáveis incluem, por exemplo, a distância até as principais atrações turísticas, a proximidade de estações de transporte público e a topografia do terreno. Por fim, a técnica MOORA é utilizada para avaliar a adequação de cada local para uma estação de bicicletas compartilhadas. Salienta-se que a MOORA é uma técnica de análise multicritério que permite a comparação de alternativas com base em múltiplos critérios (BRAUERS; ZAVADSKAS, 2010). Nesse caso, a MOORA é usada para classificar os locais de acordo com sua adequação para uma estação de bicicletas compartilhadas. A combinação dessas três técnicas de análise permite uma avaliação abrangente da adequação do local para uma estação de bicicletas compartilhadas, levando em consideração múltiplos critérios e variáveis espaciais. Ao final da aplicação da metodologia proposta no documento, é gerado um ranking dos locais mais adequados para a instalação de estações de bicicletas compartilhadas. Esse ranking é baseado na análise de múltiplos critérios, que incluem a proximidade de atrações turísticas, a disponibilidade de transporte público e a topografia do terreno. Além disso, o documento também apresenta mapas gerados pelo SIG que mostram a distribuição espacial dos locais mais adequados para a instalação de estações de bicicletas compartilhadas. Esses mapas podem ser usados para orientar a tomada de decisão sobre a localização das estações de bicicletas compartilhadas em uma outra cidade.

Em (RAUF; ALI; AL-ANSARI, 2023), os autores discutem a adequação do armazenamento de energia térmica em aquíferos na sub-bacia de Halabja-Khurmal utilizando o método AHP baseado em um SIG. Os autores utilizam seis critérios para avaliar a adequação do armazenamento de energia térmica em aquíferos na sub-bacia de Halabja-Khurmal. Cada critério é avaliado e ponderado com base em sua importância relativa para o armazenamento de energia térmica em aquíferos na sub-bacia através do método AHP. Em seguida, cada critério é dividido em subcritérios e atributos que são avaliados usando uma escala de pontuação de 0 a 10, no qual 10 indica uma condição altamente favorável para o armazenamento de energia térmica em aquíferos e 0 indica uma condição altamente desfavorável. Os subcritérios e atributos são avaliados com base em uma combinação de decisões de especialistas e informações de literatura. A pontuação para cada subcritério e atributo é ponderada com base em sua importância relativa e os resultados são combinados para gerar um mapa de adequação. O mapa de adequação mostra as áreas da sub-bacia, altamente adequadas, adequadas, fracamente adequadas ou inadequadas para o armazenamento de energia térmica em aquíferos.

Em (AJRINA *et al.*, 2020), os autores utilizam SIG e técnicas MCDM para mapear o local potencial para áreas de mineração na Indonésia. Eles utilizam informações sobre fatores naturais, ambientais e estéticos como critérios para escolha das áreas de mineração. A ponderação dos critérios utilizados pelos autores é baseada no método Fuzzy AHP. Após a definição dos pesos de cada critério, os autores utilizam três métodos: *Simple Additive Weighting (SAW)*, MOORA e *Complex Proportional Assessment (COPRAS)*. Sendo que o método SAW foi utilizado para calcular a pontuação de cada subcritério. Esseé baseado na soma ponderada dos subcritérios, em que cada subcritério é multiplicado pelo seu peso e, em seguida, somado aos outros subcritérios. O resultado é uma pontuação para cada área de mineração. Já o método MOORA foi utilizado para classificar as áreas de mineração com base nas pontuações dos subcritérios. Esseé baseado na otimização de múltiplos objetivos, em que cada objetivo é maximizado ou minimizado. O método MOORA, por sua vez, utiliza uma relação de razão para normalizar as pontuações dos subcritérios e, posteriormente, calcular a pontuação final para cada área de mineração. Por fim, o método COPRAS foi utilizado para classificar as áreas de mineração com base nas pontuações dos subcritérios. Esse é baseado na avaliação proporcional complexa, em que cada subcritério é analisado em termos de sua importância e utilidade. Tal método também utiliza uma matriz de decisão para calcular a pontuação final para cada área de mineração. Após a aplicação dos métodos MCDM, as áreas de mineração foram classificadas com base nas pontuações dos subcritérios. As áreas com as pontuações mais altas foram consideradas as mais adequadas para a mineração de areia e cascalho natural. A escolha da área mais adequada foi feita comparando as pontuações obtidas pelos três métodos e selecionando a área com a pontuação mais alta. Ao final, é gerado um mapa que mostra as áreas potenciais para mineração de areia e cascalho natural na regência de Kediri, na Indonésia.

3.3 Considerações finais

Nesse capítulo, descreveram-se os trabalhos utilizados como base para o desenvolvimento dessa tese, além do processo de seleção dos trabalhos relacionados à proposta vigente, no presente trabalho, fundamentado na metodologia de mapeamento sistemático da literatura. No próximo capítulo, será apresentada a metodologia utilizada para o desenvolvimento dos objetivos dessa tese.

CAPÍTULO

4

METODOLOGIA

4.1 Considerações iniciais

A tomada de decisão para problemas complexos com base em fontes de dados heterogêneas e múltiplas requer a estruturação de informações com representatividade adequada ao fenômeno em análise. Dimensões como temporal e espacial adicionam complexidade ao processamento de dados, extração de informações e interpretação de resultados (MORAES; NOGUEIRA; SOUSA, 2014; MALCZEWSKI; RINNER, 2015).

Liu e Zhu (2021) destaca dois grandes desafios relacionados a essas fontes de dados:

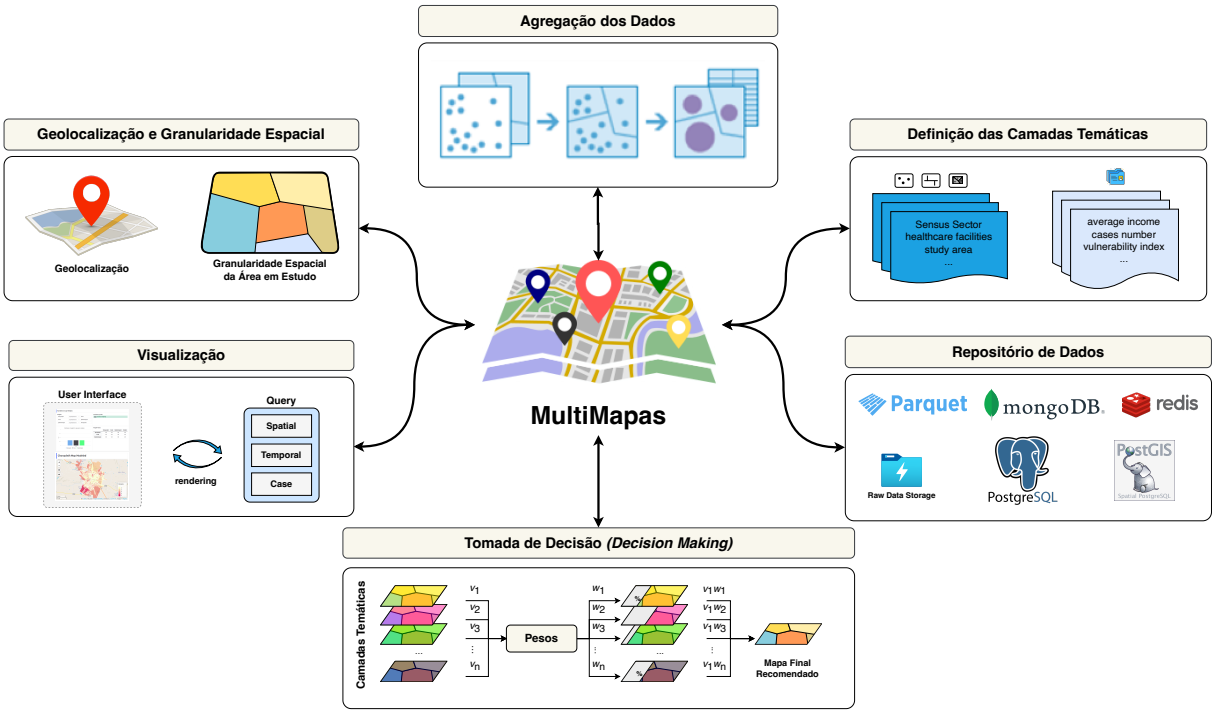
1. O acesso aos dados, já que alguns deles são confidenciais, sensíveis ou falta de profissionais para coletá-los e pré-processá-los antes da publicação;

2. Consistência entre eles em termos de granularidade tempo-espaço, quando os dados estão disponíveis, correspondem a relatórios de períodos e regiões relativamente grandes, com representantes inadequados para as previsões.

O MultiMapas, Figura 30, é uma proposta que visa superar essas barreiras. Ele pode lidar com dados de acesso aberto (de forma anônima), usar fontes diferentes (fornecidas por várias organizações) e trabalhar com granularidades espaciais distintas, independente do tipo de dado.

O MultiMapas permite que analistas e tomadores de decisão pensem sobre relacionamentos espaciais de uma forma mais sofisticada e significativa do que seria possível de outra forma, integrando de forma automática múltiplas fontes de dados e proporcionando a profissionais de diversas áreas a inspeção visual dos mapas georreferenciados múltiplos, visando realizar adaptações de acordo com seus conhecimentos.

Figura 30 – Framework MultiMapas.



Fonte: Elaborada pelo autor.

4.2 MultiMapas

O MultiMapas foi desenvolvido para auxiliar a captura, o gerenciamento e a integração de dados geoespaciais, independente do tipo do dado (vetorial ou *raster*) e da semântica.

Na Figura 30 é descrita a organização dos módulos implementados pelo MultiMapas que é composto por seis, resumidos em:

- **Módulo de Geolocalização e Granularidades Espaciais:** Este módulo é responsável por carregar, limpar, anonimizar e geolocalizar dados de múltiplas fontes. Essa camada também define a granularidade espacial de interesse relacionada ao fenômeno em estudo.

- **Módulo de Agregação de Dados:** Pretende realizar a agregação de dados, considerando a granularidade espacial da área de estudo.

- **Módulo de Definição de Camadas Temáticas:** Realiza a definição e seleção de camadas temáticas (dados agregados) com propriedades de dados e representação espacial que sintetizam o fenômeno em estudo.

- **Módulo Tomada de Decisão (Decision Making):** Módulo responsável pelo gerenciamento dos algoritmos utilizados na integração e determinação da influência de cada camada temática para a tomada de decisão, considerando o padrão espacial do fenômeno.

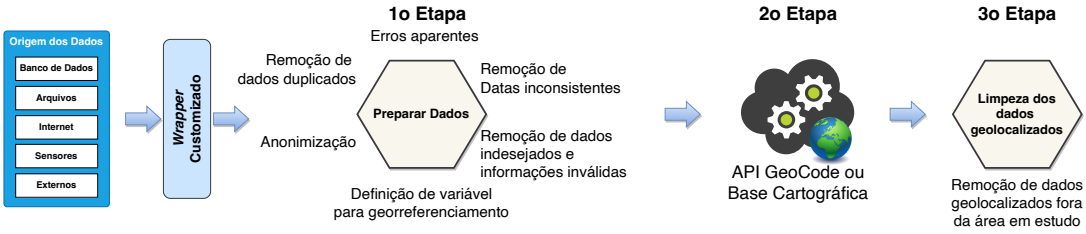
Os algoritmos utilizados, nessa camada, devem ser capazes de tomar decisões levando em conta múltiplos objetivos conflitantes.

- **Módulo de Visualização:** Camada de visualização dos dados e do mapa temático gerado pelo módulo tomada de decisão através da *interface web*.
- **Módulo Repositório de Dados:** A principal finalidade deste módulo é armazenar e fornecer todos os dados geoespaciais.

4.2.1 Módulo de Geolocalização e Granularidades Espaciais

O processo de Geolocalização proposto e realizado pelo Módulo de Geolocalização e Granularidades Espaciais, conforme pode ser observado na Figura 31, é realizado em quatro etapas: (i) A partir da fonte de dados, é definido um *wrapper* customizado que encapsula todas as propriedades daquela fonte de dados e os atributos que serão utilizados como identificadores de localização para o georreferenciamento; (ii) remove todos os dados sensíveis, duplicados e inconsistentes, além de informações indesejadas e inválidas; (iii) a partir do conjunto de variáveis definidas pelo *wrapper*, utiliza-se uma API, pública ou privada, de geolocalização ou uma base cartográfica para geolocalizar os dados; (iv) após a geolocalização dos dados, é realizada uma verificação dos dados geolocalizados considerando a área do fenômeno em estudo, na qual são desconsiderados dados geolocalizados fora dela, sendo essa etapa chamada de Limpeza de Dados de Geolocalização. As transformações e os resultados de cada etapa são armazenados no Módulo de Repositório de Dados para todo o processo possa ser revisado e auditado.

Figura 31 – Workflow de geolocalizado implementado no MultiMapas Framework.



Fonte: Elaborada pelo autor.

Nesse modelo também é responsável pela definição da área objeto de estudo e da granularidade espacial a ser considerada nos demais módulos do MultiMapas

4.2.2 Módulo de Agregação de Dados

Esse módulo é responsável pela agregação dos dados geolocalizados no Módulo de Geolocalização na granularidade espacial defina na área objeto de estudo, por meio de uma função de *de join spatial* que realiza o cruzamento de cada dado geolocalizado na granularidade espacial da área objeto de estudo.

Nesta etapa garante que os dados geolocalizados e a granularidade estejam em um sistema de referência espacial comum. Isso envolve a transformação dos dados para um único sistema de coordenadas, como latitude e longitude (por exemplo, WGS84), em um sistema de projeção específico que pode ser o mesmo da granularidade espacial da área objeto de estudo ou outro sistema de projeção.

4.2.3 Módulo de Definição das Camadas Temáticas

Em SIG, as camadas temáticas referem-se às diferentes fontes de dados geográficos que são organizadas e exibidas em um mapa. Cada camada temática representa uma categoria específica de informações geográficas (O'SULLIVAN; UNWIN, 2003). Segundo Longley *et al.* (2015) uma camada temática é uma representação visual e geográfica de um determinado conjunto de dados relacionados a um tema específico.

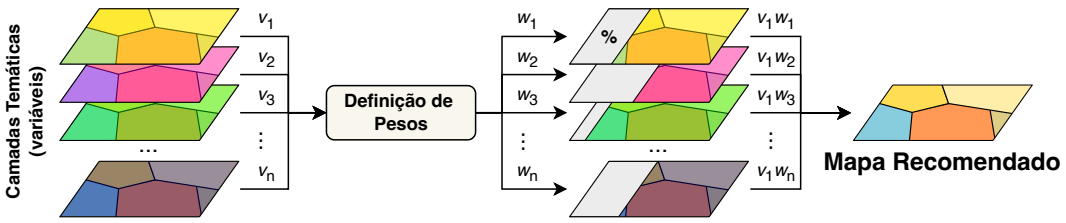
As camadas temáticas são elementos fundamentais em SIG, permitindo a visualização, análise e tomada de decisões com base em informações geográficas. Através da combinação de várias camadas temáticas, é possível criar mapas temáticos ricos e explorar as relações espaciais dos dados (O'SULLIVAN; UNWIN, 2003; LONGLEY *et al.*, 2015).

Esse módulo é responsável pela definição e seleção das camadas temáticas (dados agregados) com a representação espacial que sintetizam o fenômeno em estudo e que serão utilizadas para a composição do mapa global, definido no Módulo de Tomada de Decisão (*Decision Making*).

4.2.4 Módulo de Tomada de Decisão

O processo de tomada de decisão envolvendo dados espaciais pode ser pensado como um processo que combina e transforma uma série de dados geográficos, camadas temáticas de entrada, por meio de fatores de ponderação, pesos, que indicam o grau de importância de cada dado geográfico sobre os demais, em uma decisão resultante, saída, veja a Figura 32. O resultado é uma agregação de informações multidimensionais em um único mapa global de saída: a decisão. Esse processo inclui, além dos dados geográficos, as referências do decisor e a manipulação de dados e preferências de acordo com regras de decisão especificadas.

Figura 32 – A estrutura da tomada de decisão de múltiplos atributos baseada em GIS.



Fonte: Elaborada pelo autor.

Um dos pontos-chave no processo de tomada de decisão com dados espaciais é encontrar uma atribuição de pesos para as camadas temáticas (critérios ou variáveis) que equilibre a visão do usuário e a relação ótima entre as camadas temáticas do ponto de vista espacial. Os pesos, $w_1, w_2, \ldots, w_n$, são normalmente assumidos para atender às seguintes condições: $w_n \in [0,1]$, e $\sum_{i=1}^{n} w_i = 1$.

O MultiMaps implementa duas abordagens para definição de pesos para cada camada temática, descritas a seguir:

1. Através do método AHP a partir do módulo de visualização, no qual o usuário define o grau de importância da camada temática por meio de comparações pareadas com base na escala fundamental do método;

2. Mediante um algoritmo evolutivo, o GIS-moGA, que otimiza a dependência e a heterogeneidade espacial das camadas temáticas.

Após definir o grau de importância de cada camada temática, aplica-se uma regra de decisão compensatória, cujo objetivo é reunir todas as camadas temáticas em uma única camada temática de adequação. MultiMapas usa combinação linear ponderada (WLC), na qual a camada temática de adequação final é derivada multiplicando cada camada temática por seu peso relativo seguido pela soma dos resultados, conforme a Equação 4.1.

$$\mu_i = \sum_{i=1}^{n} w_i \times v_i \tag{4.1}$$

onde v_i é o i-ésimo mapa temático e w_i é o i-ésimo fator de ponderação do grau de importância do i-ésimo mapa temático da área objeto de estudo.

Algoritmo Genético Multi-objetivo GIS-moGA

Segundo, Almeida (2012) a dependência espacial e a heterogeneidade espacial estão imbricadas, conduzindo, por consequência, às dificuldades de especificações de modelos espaciais mais apropriados. Dessa forma, é proposto algoritmo genético multi-objetivo, o GIS-moGA, que utiliza a maximização do índice I de Moran e a minimização do índice local de Moran LISA para produzir um *trade-off* de Pareto entre a dependência e a heterogeneidade espacial, respectivamente.

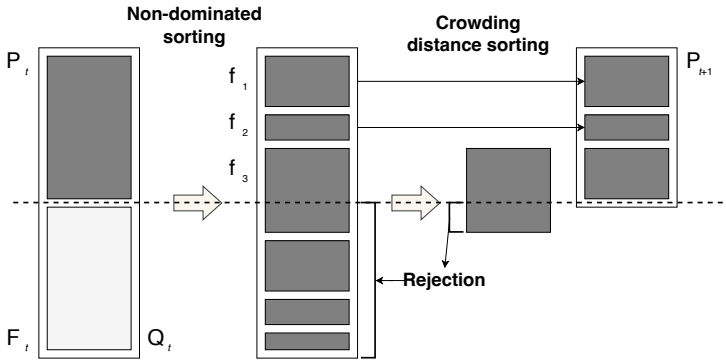
O GIS-moGA é baseado no NSGA-II (*Non-dominated Sorting Genetic Algorithm II (NSGA-II)*) (DEB *et al.*, 2002), um dos algoritmos evolutivos multi-objetivos (*Multi-objective Evolutionary Algorithms (MOEA)*) mais utilizados (GASPAR-CUNHA; TAKAHASHI; ANTUNES, 2012; MALCZEWSKI; RINNER, 2015; UYDURAN *et al.*, 2016; BEIRIGO; SANTOS, 2016). O NSGA-II envolve dois aspectos: ordenação rápida não-dominada (*non-dominated sorting*) de indivíduos e a estratégia de seleção elitista (*crowding-distance sorting*). A ordenação

não-dominada é formada por índices de classificação não-dominados pela distância de aglomeração. A relação de dominância de Pareto determina a classificação das soluções não-dominadas. Considerando um problema multiobjetivo com duas funções, a dominação de Pareto é definida a seguir. Para vetores de soluções x_1 e x_2 é chamada de solução dominante de Pareto quando as duas condições seguintes foram atendidas:

$$[f_i(x_1) \leq f_i(x_2), i = 1,2] \& [f_i(x_1) \leq f_i(x_2), i \in \{1,2\}] \tag{4.2}$$

O processo de evolução no NSGA-II é ilustrado na Figura 33. Os indivíduos partem de dois grupos na n-ésima geração: indivíduo pai (P_t) e indivíduos descendentes (F_t). A classificação não-dominada D e o índice de distância de aglomeração para cada indivíduo são calculados para a nova população de indivíduos $Q_t = (P_n, Q_n)$. O índice de distância de superposição representa a uniformidade da distribuição individual através da população atual. Os melhores são selecionados com base na classificação não-dominada. Para indivíduos com o mesmo valor de classificação, o índice de distância de aglomeração é usado para desempate, em que se opta por aqueles com índice mais alto. Esse processo na n-ésima geração termina quando a nova população P_{t+1} atinge o número total de indivíduos.

Figura 33 – Processo de evolução do algoritmo NSGA-II.



Fonte: Deb *et al.* (2002).

Algoritmo 4 descreve a estrutura do GIS-moGA.

Cada indivíduo representa um mapa temático global da *k*ésima área de estudo com pesos em que a soma é igual a um. O algoritmo começa com uma população (p_t), linha 2, com um conjunto de pesos aleatórios w_i, com $i = 1 \ldots n$, no qual n é o número de variáveis utilizadas para compor o mapa temático global μ. Após gerar a população inicial, cada indivíduo é avaliado, calculando-se o I de Moran Global e a variância do Índice de Moran Local (LISA). Algoritmo 5 descreve como a função de avaliação é usada no GIS-moGA proposto.

A função *selection*, linhas 6 do Algoritmo 4, aplica o método *fast-non-dominated-sort* para ordenar as soluções não-dominadas em duas etapas. Primeiro, para todas as soluções, um grau de dominância (n_p) é calculado com base no número de soluções que dominam uma solução

Algoritmo 4 – Estrutura do GIS-moGA

1: $t \leftarrow 0$
2: initialize p_t
3: evaluate p_t
4: **enquanto** not stop-criterion **faça**
5: $t \leftarrow t+1$
6: $S_t \leftarrow$ select p_{t-1}
7: $O_t \leftarrow$ reproduce S_t
8: $X_t \leftarrow$ evaluate O_t
9: $p_t \leftarrow S_t \cup X_t$
10: **fim enquanto**

Algoritmo 5 – Algoritmo da função de avaliação do GIS-moGA

1: **função** GLOBAL_THEMATIC_MAP(p_i)
2: $\mu_i \leftarrow \sum_i w_i \times v_i$ ▷ apply WLC from w_i
3: **retorna** μ_i
4: **fim função**
5: **função** EVALUATE(p_t)
6: **para** $p_i \in p_t$ **faça**
7: $\mu_i \leftarrow global\_thematic\_map(p_i)$
8: $I_i \leftarrow moran\_global(\mu_i)$
9: $LISA_i \leftarrow lisa(\mu_i)$
10: fitness[i] $\leftarrow (I_i, variance(LISA_i))$
11: **fim para**
12: **retorna** p_t evalueted
13: **fim função**

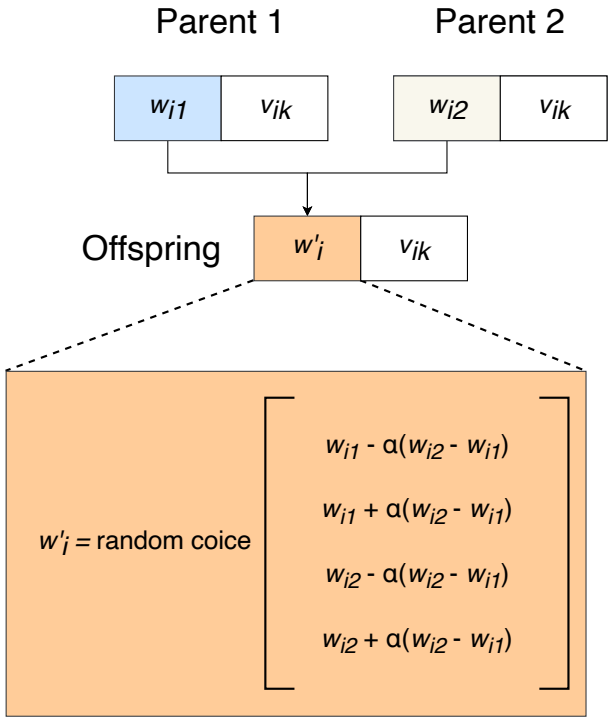
$s = (I$ de Moran Global, variância($LISA$)). Se o valor de n_p for 0, uma solução s não é dominada e fará parte do primeiro conjunto. O segundo passo é separar as soluções em grupos s_s em ordem de dominância. Assim, cada indivíduo adicionado a um conjunto s_s é retirado da população e os indivíduos dominados por ela têm seu valor de n_p diminuído. A segunda etapa é repetida até que não haja mais indivíduos na população.

Após indexar as soluções, dentro dos conjuntos de soluções não-dominadas, as soluções são ordenadas pela distância de seus valores de aptidão (*fitness*).

Uma vez selecionada a população S_t, linha 6, essa população é utilizada para a reprodução dos descendente, linha 7. Primeiramente, dois pais (p_1 e p_2) são selecionados, da seguinte forma: seleciona um indivíduo $S_t[i]$ e seu próximo $S_t[i+1]$, se $S_t[i]$ for o último indivíduo, considera-se o primeiro indivíduo de S_t. Dos pais selecionados, um conjunto de descendentes é gerado aplicando-se um descendente predador, neste caso, um indivíduo com pesos aleatórios, se a taxa de mutação for satisfeita, o que significa gerar aleatoriamente $\lambda \in \{0,1\}$ com $\lambda < mutRate$ ou o operador *crossover* caso contrário. O operador de *crossover* implementado pelo GIS-moGA é o *blend alpha crossover* (blx-α) baseado em (HAUPT; HAUPT, 2004). A Figura 34 ilustra o procedimento de *crossover* utilizado pelo GIS-moGA a partir de dois pais, p_1 e p_2. O Algoritmo 6

descreve o funcionamento da função de reprodução usada pelo GIS-moGA.

Figura 34 – Blx- α crossover utilizado pelo GIS-moGA.



Fonte: Elaborada pelo autor.

Após gerar e avaliar o conjunto de descendentes, uma nova população (p_t) é gerada a partir da população selecionada (S_t) e dos descendentes avaliados (X_i).

4.2.5 Módulo de Repositório de Dados

O repositório de Dados concentra todas as informações dos dados e suas transformações, as granularidades espaciais da área objeto de estudo a serem utilizadas, além das informações utilizadas no Módulo de Tomada de Decisão. A arquitetura do Repositório de Dados consiste em: (i) um *data lake* [1] contendo três camadas (*bronze*, *silver* e *gold*); (ii) um banco de dados relacional com extensão espacial (PostgreSQL 12 com a extensão PostGIS); (iii) dois bancos NoSQL para armazenamento de informações de contexto (MongoDB) e informações voláteis (Redis), que não precisam ser persistidas, para melhorar a usabilidade do módulo de visualização.

A Figura 35 apresenta a jornada do dado a ser geolocalizado dentro da arquitetura do MultiMapas, da ingestão até o Módulo de Tomada de Decisão.

[1] Um *data lake* refere-se a um repositório de armazenamento massivamente escalável que contém uma grande quantidade de dados brutos em seu formato nativo até que seja necessário, além de sistemas de processamento (mecanismo) que podem ingerir dados sem comprometer a estrutura de dados (MILOSLAVSKAYA; TOLSTOY, 2016).

Algoritmo 6 – Algoritmo da reprodução utilizado no GIS-moGA.

```
1:  função REPRODUCE(S_t)
2:      i ← 0
3:      n_pop ← len(S_t)                          ▷ selected population size
4:      offspring ← empty list
5:      enquanto i < n_pop faça
6:          λ ∈ {0,1}
7:          se i < (n_pop − 1) então
8:              p_1 ← S_t[i]                      ▷ parent 1
9:              p_2 ← S_t[i+1]                    ▷ parent 2
10:         senão
11:             p_1 ← S_t[i]                      ▷ parent 1
12:             p_2 ← S_t[0]                      ▷ parent 2
13:         fim se
14:         se λ < mutRate então
15:             offspring[i] ← predator individual
16:         senão
17:             offspring[i] ← crossover(p_1, p_2)
18:         fim se
19:         i ← i+1
20:     fim enquanto
21:     retorna offspring
22: fim função
```

4.2.6 Visualização

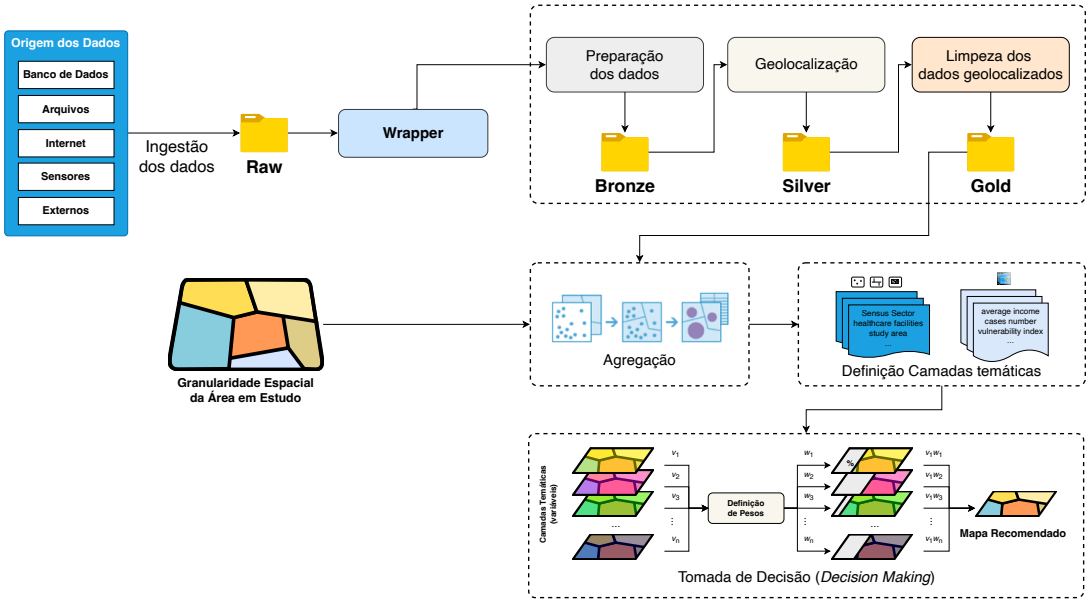
Após a ingestão dos dados, agregação e definição das camadas temáticas (Figura 35) com suas respectivas propriedades, o usuário, via interface web, utiliza uma técnica MCDM para definir a influência de cada camada temática, produzindo um mapa coroplético final com o padrão espacial do fenômeno em estudo, conforme Figura 36.

O *frontend* do MultiMapas foi criado usando uma combinação de JavaScript, HTML e CSS e mapas interativos usando a biblioteca Leaflet versão 1.8.0. No *backend* utiliza-se Python como linguagem de programação e o *framework* Django na versão 3.1.7. Na Figura 37, é apresentada a Arquitetura conceitual do MultiMapas na qual é evidenciado as tecnologias utilizadas.

Na Figura 38, é apresentada a tela inicial da interface *web* do MultiMapas em que o usuário pode realizar a inspeção visual das informações já inseridas ou, então, realizar a análise e fusão das camadas temáticas já definidas através do método AHP ou o algoritmo evolutivo.

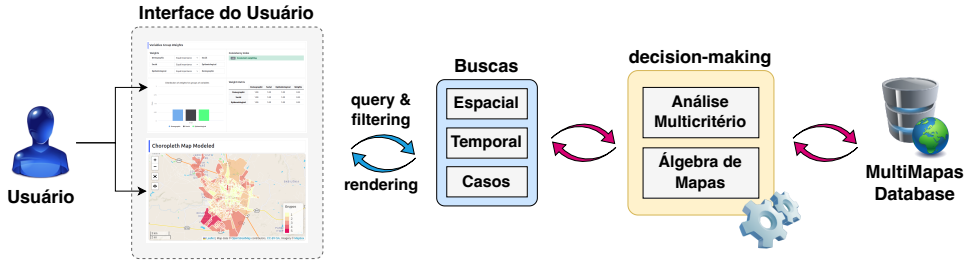
A Figura 39 é apresentada à tela com um exemplo no qual o usuário define a contribuição de cada camada temática através do método AHP e a Figura 40 um exemplo em que o usuário, após definir a contribuição de cada camada temática, pode então comparar seu mapa modelo com a visão de um especialista.

Figura 35 – Jornada dos dados na arquitetura do MultiMapas da ingestão até o módulo *decision-making*.



Fonte: Elaborada pelo autor.

Figura 36 – Visão geral do mecanismo de inspeção visual.



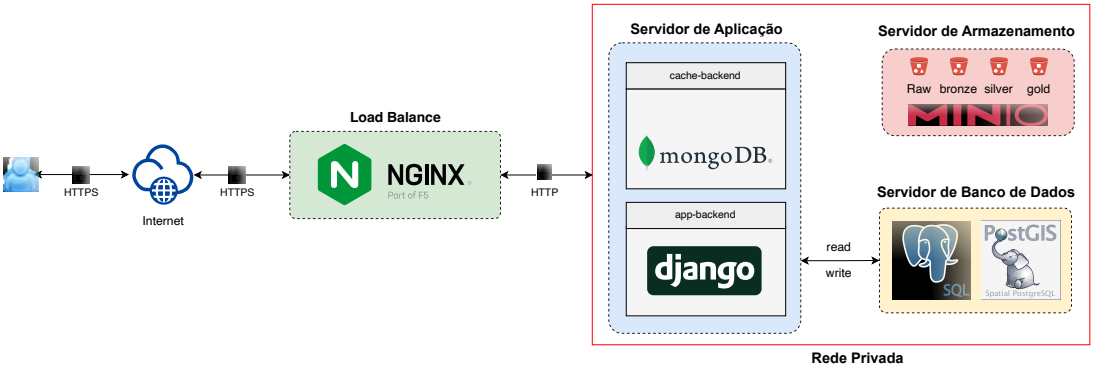
Fonte: Elaborada pelo autor.

A Figura 41 mostra um exemplo de *heatmap* (mapa de calor) com os casos de um agravo de saúdo, permitindo uma inspeção visual das áreas que concentram o maior número desses casos, podendo o usuário filtrar os casos por data de notificação e semana epidemiológica. A Figura 42 mostra a evolução dos casos ao longo do tempo de forma acumulada ou com janelas deslizantes, aonde o usuário pode acompanhar a evolução da distribuição dos agravos de saúde ao longo do tempo.

4.3 Considerações Finais

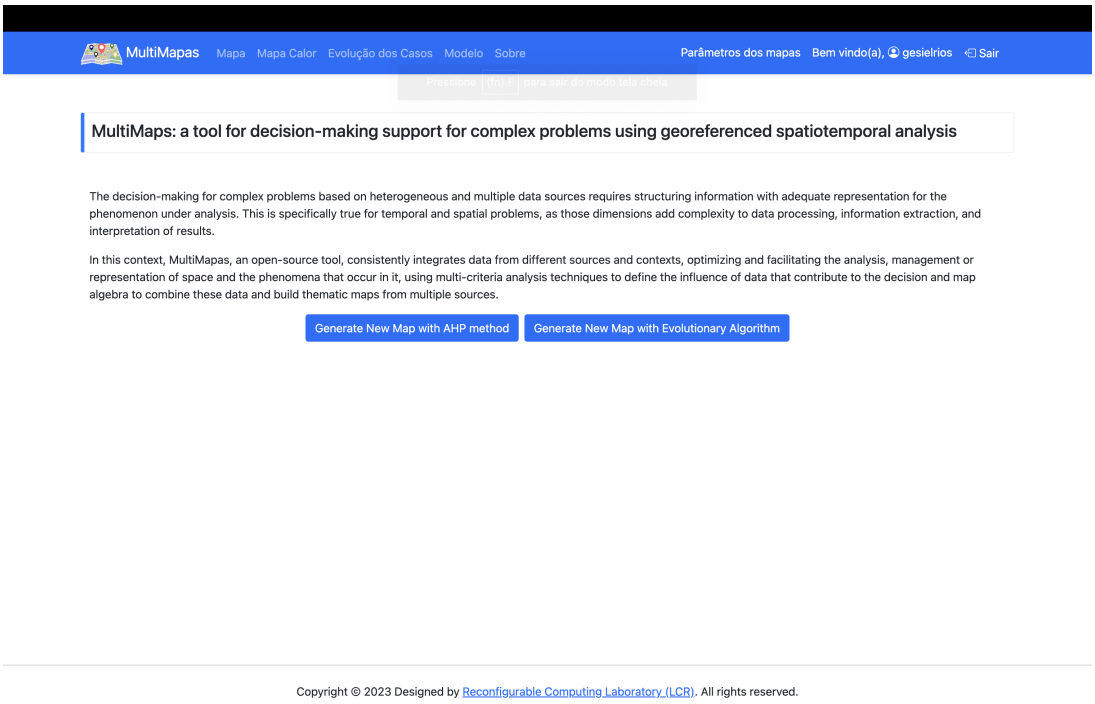
Esse capítulo apresentou o desenvolvimento do MultiMapas, um *framework* para a manipulação de dados com características espaciais de múltiplas e heterogêneas fontes, a implementação de uma interface amigável para visualização e interação dessas fontes de dados, além da implementação de um método multicritério e um algoritmo genético multiobjetivo que otimiza a dependência e heterogeneidade espacial. No próximo capítulo, são descritos os estudos

Figura 37 – Arquitetura conceitual do MultiMapas.



Fonte: Elaborada pelo autor.

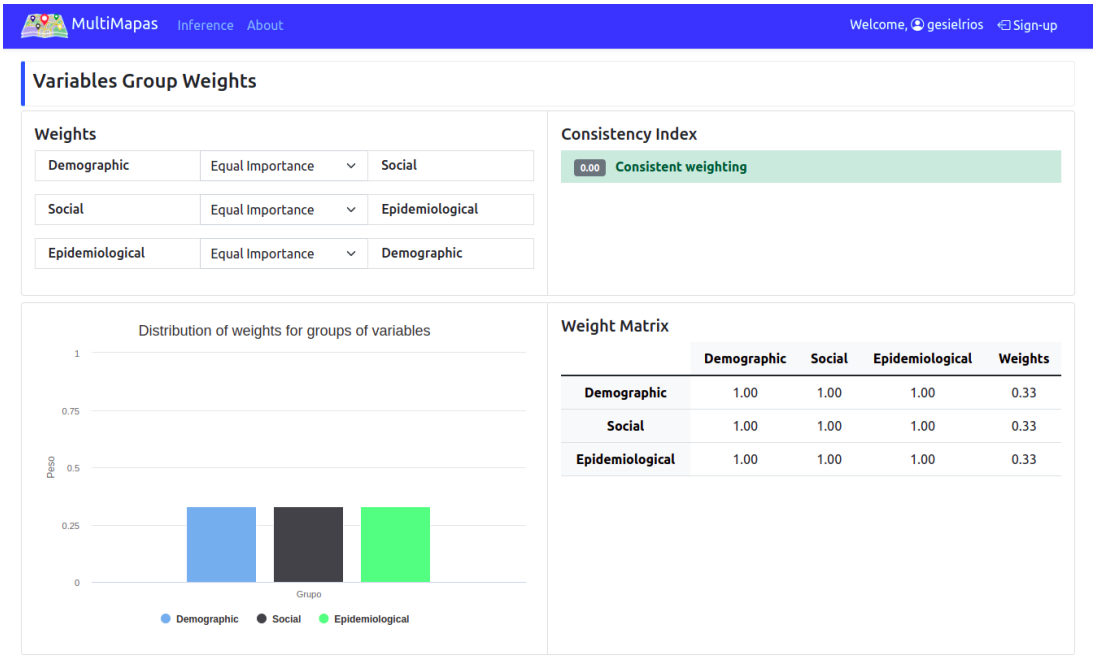
Figura 38 – Tela inicial do MultiMapas.



Fonte: Elaborada pelo autor.

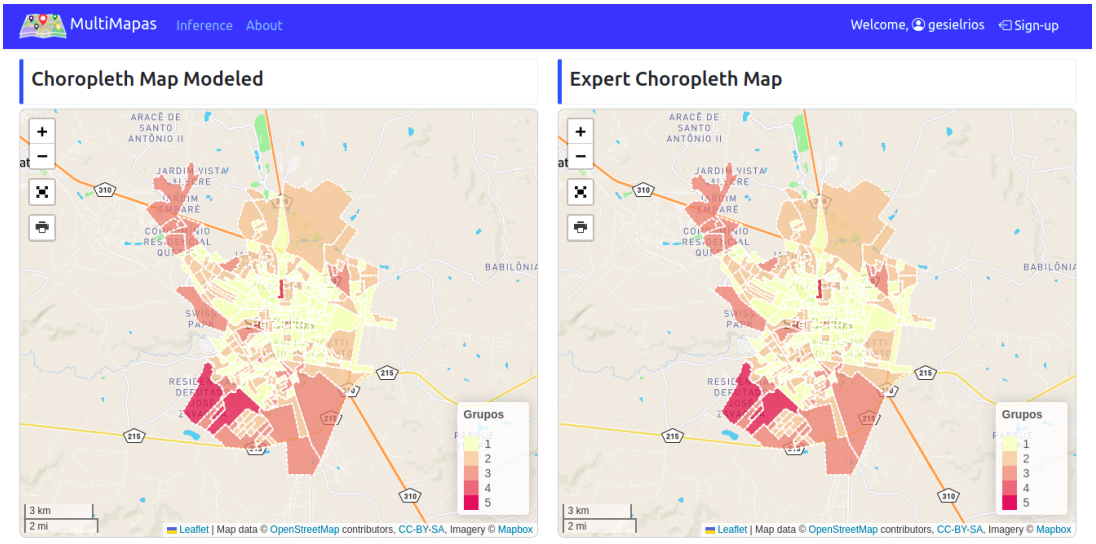
exploratórios utilizados para validar a metodologia proposta ness tese, descrevendo por meio de três estudos de casos.

Figura 39 – Contribuição hierárquica de cada camada temática por meio do método AHP.



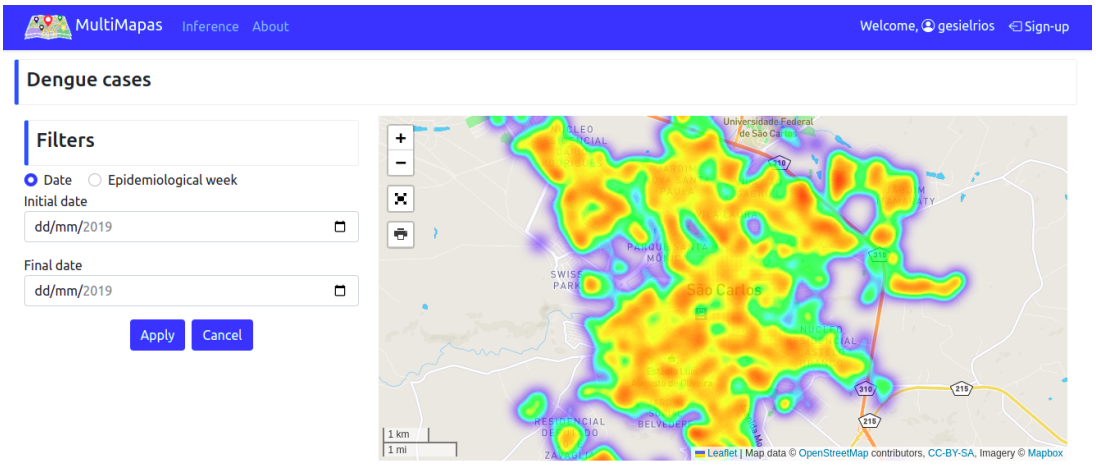
Fonte: Elaborada pelo autor.

Figura 40 – Mapa coroplético modelado pelo usuário e pelo especialista em saúde.



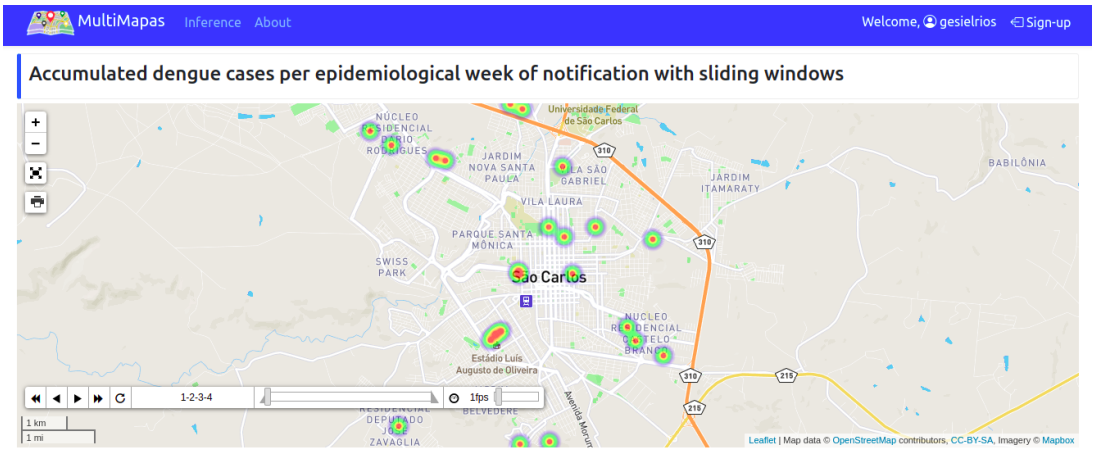
Fonte: Elaborada pelo autor.

Figura 41 – Exemplo de *Heatmap* de um agravo de saúde.



Fonte: Elaborada pelo autor.

Figura 42 – Exemplo de *Heatmap* ao longo do tempo.



Fonte: Elaborada pelo autor.

CAPÍTULO

5

RESULTADOS

5.1 Considerações iniciais

A tomada de decisão para problemas complexos com base em fontes de dados heterogêneas e múltiplas requer a estruturação de informações com representação adequada ao fenômeno em análise. Este capítulo apresenta os resultados computacionais do MultiMaps, em três estudos de casos: (i) área de saúde: controle de múltiplos agravos de saúde; (ii) seguro agrícola para produção de tomate; e (iii) aplicação da estatística de varredura. Os *insights* obtidos permitem ao decisor entender melhor tanto as entradas de dados quanto os diferentes resultados possíveis do modelo multicritério de tomada de decisão, considerando diferentes combinações de pesos.

5.2 Estudo de caso 1: Múltiplos agravos de saúde

O estudo dos padrões de distribuição geográfica das doenças e suas relações com fatores socioambientais constitui-se no objeto do que, hoje, chamamos de Epidemiologia Geográfica. Compreender e mapear os riscos e dificuldades de enfrentamento das doenças considerando esses fatores socioambientais é complexo e envolve variáveis com características diversas e de origem difusa (ANDRADE *et al.*, 2007).

Para estabelecer políticas coerentes, o território (ou espaço geográfico) é parte fundamental no planejamento de ações de promoção e atenção integral à saúde, pois é dele que se obtêm informações de distribuição demográfica e epidemiológica envolvendo o contexto social, político, cultural e administrativo (LIMA *et al.*, 2019; MELO; MELO; MORAES, 2022).

Uma das questões para a tomada de decisão na área da saúde é a integração consistente de informações de diferentes fontes (LIU; ZHU, 2021). Por exemplo, a Vigilância em Saúde lida com dados de pacientes em diferentes níveis de atenção por unidades de saúde, vetores de transmissão, aspectos socioeconômicos, multimorbidades (que podem gerar confusão diagnóstica

e agravos), bem como os recursos do Sistema Único de Saúde distribuídos nas regiões da cidade.

A tomada de decisão também requer a estruturação de informações com representação adequada do fenômeno sob análise. As informações devem, por exemplo, ser agregadas segundo uma granularidade espacial específica que pode diferir entre regiões de dados sociais e epidemiológicos e mesmo entre dados de diferentes agravos de saúde. A granularidade do tempo é crucial, pois os ciclos epidemiológicos podem diferir, além dos aspectos sazonais (LIMA *et al.*, 2019; MELO; MELO; MORAES, 2022).

5.2.1 Área de estudo

O estudo de caso foi realizado no município de São Carlos (Figura 43), cidade de médio porte do interior do estado de São Paulo, Sudeste do Brasil. Localizada na região centro-leste do estado, especificamente nas coordenadas $22^o1'4''$ latitude Sul e $47^o53'27''$ latitude Oeste, São Carlos possuía uma área territorial total de 1.136,907 km^2, altitude média de 856 metros, densidade demográfica de 195,15 habitantes/km^2 e população residente de 221.950 habitantes em 2010. No que concerne aos aspectos socioeconômicos, o município apresentava um índice de Gini de 0,63, Índice de Desenvolvimento Humano (IDH) de 0,805 e produto interno bruto de R$ 6.712.498,00 para o mesmo ano de 2010. A granularidade espacial adotada foi a dos setores censitários[1] da área urbana do município.

Figura 43 – Município de São Carlos-SP e seu perímetro urbano.

Fonte: Elaborada pelo autor.

[1] O setor censitário é a unidade territorial estabelecida para fins de controle cadastral, formado por área contínua, situada em um único quadro urbano ou rural, com dimensão e número de domicílios que permitam o levantamento por um recenseador do Instituto Brasileiro de Geografia e Estatística (IBGE) (CENSO, 2010).

5.2.2 Seleção das variáveis

Para sintetizar as informações relevantes à tomada de decisão na área da saúde, foi consultado um grupo de especialistas multidisciplinares com a presença de profissionais médicos, enfermeiros, epidemiologistas, estatísticos, especialistas em geoprocessamento, entre outros, para definir um conjunto de possíveis variáveis de entrada que um sistema de modelagem precisaria para realizar o mapeamento de regiões críticas em relação a epidemias.

A partir de entrevistas com um grupo multidisciplinar de especialistas e disponibilidade de dados, foi mapeado um conjunto de oito variáveis, agrupadas em três grupos segundo aspectos demográficos, sociais e epidemiológicos. Três variáveis trazem aspectos demográficos, uma variável considera critérios de vulnerabilidade social e quatro variáveis epidemiológicas contêm a contagem de casos confirmados de três epidemias com características e contágios diferentes (dengue, covid-19 e tuberculose) fornecidas pela Vigilância Epidemiológica de São Carlos (VIGEP-SP) e a influência da distribuição das unidades de saúde. A Tabela 8 apresenta o conjunto de variáveis selecionadas para a modelagem.

Tabela 8 – Variáveis selecionadas para a modelagem.

Variável	**Nome Variável**	**Fonte dos Dados**
Densidade Demográfica	*DensDemoG*	Censo IBGE 2010
Média de Moradores por Domicílio	*MorPDomG*	Censo IBGE 2010
Percentual de População com idade Superior a 60 anos	*PercPop*60*G*	Censo IBGE 2010
Índice Paulista de Vulnerabilidade Social	IPVS	SEADE SP
Presença de Unidades de Saúde	*USCount*	VIGEP-SC
Contagem de casos de dengue	*DengueCount*	VIGEP-SC
Contagem de casos de COVID-19	*CovidCount*	VIGEP-SC
Contagem de casos de tuberculose	*tbCount*	VIGEP-SC

Fonte: Dados da pesquisa.

5.2.3 Descrição e Mapeamento das Variáveis

Presença de Unidades de Saúde

Um dos principais pontos de convergência das pessoas infectadas que buscam auxílio de saúde, as unidades de atendimento de saúde têm grande probabilidade de se tornarem pontos vulneráveis do ponto de vista da disseminação do contágio.

As unidades de saúde foram geolocalizadas a partir de seus endereços fornecidos pela Secretaria de Saúde, do município de São Carlos, por meio da geocodificação de endereços. Essas unidades foram classificadas conforme as suas características e capacidade de atendimento e abrangência conforme a Portaria nº. 2.488 do Ministério da Saúde (FEDERAL, 2011). Na Tabela 9, apresentam-se os grupos de classificação das unidades de saúde.

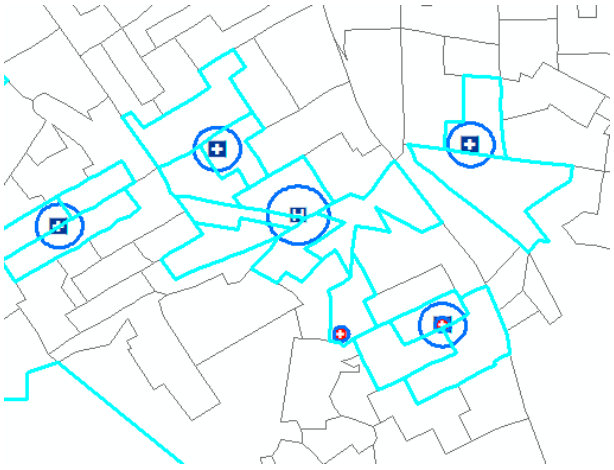
Tabela 9 – Variáveis selecionadas para a modelagem.

Grupo	**Sigla**	***Score***	**Raio de Abrangência** (*m*)
Unidade de Saúde da Família	USF	1	40
Clínicas de Convênios	CV	2	80
Unidade Básica de Saúde	UBS	3	120
Unidade de Pronto Atendimento	UPA	4	160
Hospitais	HOSP	5	200

Fonte: Dados da pesquisa.

Com a definição de um raio de abrangência e a partir dos pontos espacializados foi gerado uma camada de *buffer* que é cruzada com a camada de setores censitários. O *score* para essa variável em cada setor é calculado pela soma dos *scores* de cada unidade de saúde cujo *buffer* intersecta o referido setor. Na Figura 44 apresenta-se um recorte com a intersecção da camada dos raios de abrangência das unidades de saúde com os setores censitários.

Figura 44 – Recorte com a intersecção da camada dos raios de abrangência das unidades de saúde com os setores censitários.



Fonte: Elaborada pelo autor.

Densidade Demográfica

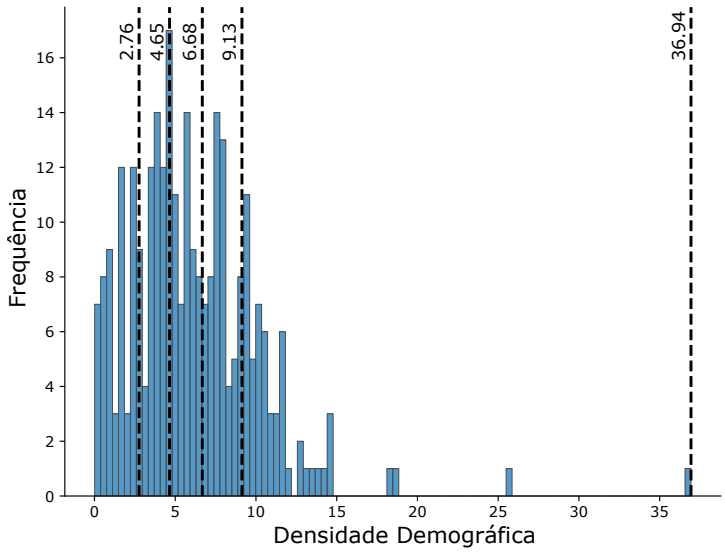
Quanto maior o número de indivíduos numa determinada região, maior a probabilidade de que os indivíduos se infectem mutuamente. Desta forma, analisar a densidade demográfica de cada setor ajuda a identificar áreas com maior capacidade de aglomeração.

Para essa análise utilizou-se a base de dados de população do censo demográfico do IBGE, realizado no ano de 2010, com agregação no nível de setores censitários. Com o total de população agregado para os setores calculou-se a densidade demográfica em habitantes por quilômetro quadrado de acordo o valor da área de cada polígono correspondente aos setores.

Uma vez em que estamos tratando de quase três centenas de setores, e devido à amplitude desses resultados, é preciso, a fim de poder ranqueá-los e atribuir-lhes *scores* para representá-los em uma classificação da variável por intervalos. Existem diferentes métodos de classificação de dados descritos na literatura. Matsumoto, Catão e Guimarães (2017) faz uma descrição de vários destes métodos destacando a importância na escolha de um adequado para a fiel representação que se deseja. Nesse estudo, em partícular, utilizou-se para classificar dados, a fim de atribuir-lhes *scores*, o método de quantis, que tem como principal característica formar classes com um número aproximado de feições atribuídas. Os valores de densidade demográfica foram então fatiados em cinco grupos conforme os seus quintis.

Na Figura 45 apresenta-se o histograma dos dados de densidades demográficas e na Figura 46 pode-se visualizar a distribuição da densidade demográfica nos setores.

Figura 45 – Histograma dos dados de densidade demográfica.



Fonte: Elaborada pelo autor.

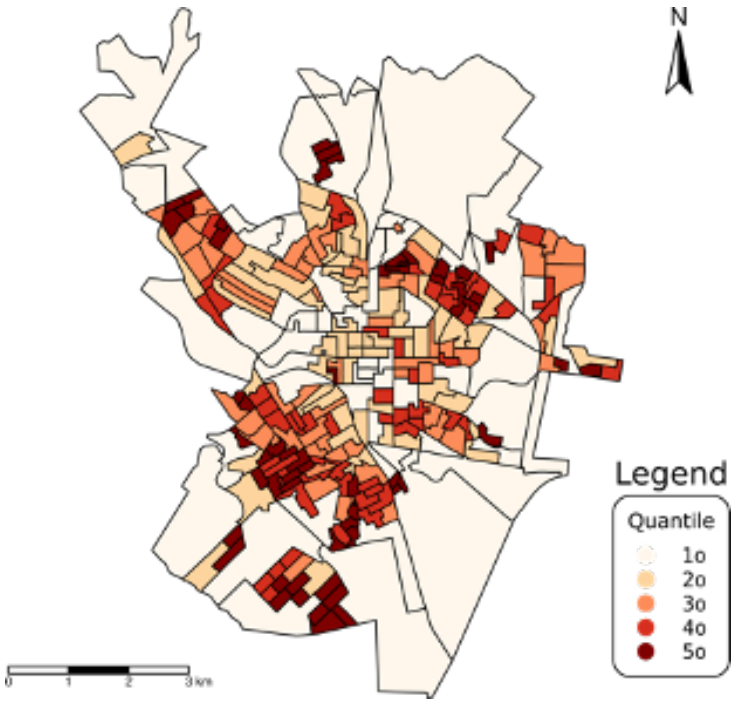
A cada grupo foi associado um *score* variando de 1 a 5 pontos conforme o aumento da densidade demográfica.

Média de Moradores por Domicílio

De forma análoga à densidade demográfica, é de se esperar que no âmbito do domicílio também tenhamos maior probabilidade de contágio com um maior número de moradores. Isto também está em acordo com o pensamento trazido à discussão pelos profissionais de saúde de que deveria ser considerado o número de pessoas dormindo em um mesmo dormitório.

Nesse caso também foi utilizado a base de dados do censo IBGE 2010. Esse dado leva em consideração a população total no setor e a respectiva contagem de domicílios. O mesmo

Figura 46 – Representação coroplética da distribuição da densidade demográfica.



Fonte: Elaborada pelo autor.

procedimento de fatiamento utilizado para a variável anterior foi adotado para essa variável.

Na Figura 47, apresenta-se o histograma dos dados de média de moradores por domicílio e já a Figura 48 pode-se visualizar a distribuição da média de moradores por domicílio nos setores segundo esses grupos.

A cada grupo foi associado um *score* variando de 1 a 5 pontos conforme o aumento da média de moradores por domicílio.

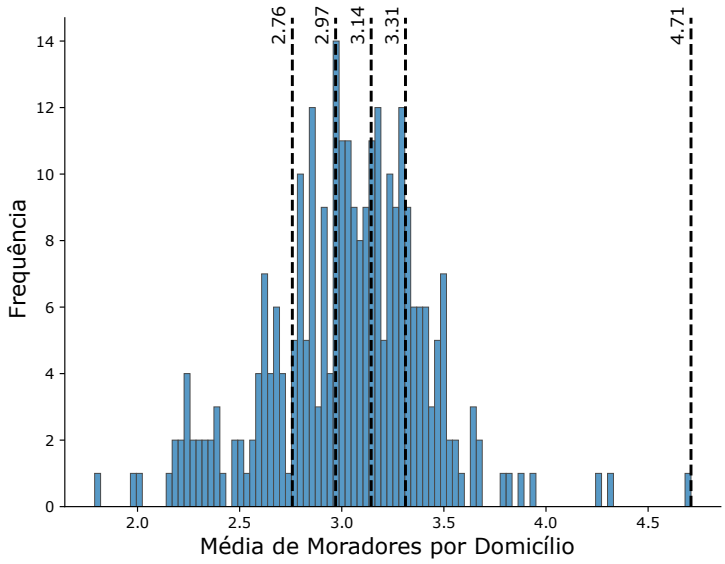
Percentual de População com Idade Superior a 60 anos

Do ponto de vista da criticidade da epidemia, é sabido que a idade é um dos fatores de risco mais relevante. A fim de contemplar, na modelagem, essa característica foi contabilizado, a partir dos dados censitários, os acumulados de população com idade acima de 60 anos e calculado o valor percentual em relação ao total, para cada setor. O mesmo procedimento de fatiamento utilizado para as duas variáveis anteriores foi adotado nesse caso.

Na Figura 49 apresenta-se o histograma dos dados de percentual de população com idade superior a 60 anos e a Figura 50 pode-se visualizar a distribuição do percentual de população com idade superior a 60 anos nos setores segundo esses grupos.

A cada grupo foi associado um *score* variando de 1 a 5 pontos conforme o aumento do percentual de população com idade superior a 60 anos.

Figura 47 – Histograma dos dados de média de moradores por domicílio.



Fonte: Elaborada pelo autor.

Índice Paulista de Vulnerabilidade Social

O conhecimento da condição prévia de vulnerabilidade social de uma região é outro fator relevante a ser considerado frente aos desafios de enfrentamento de uma epidemia, principalmente do ponto de vista da criticidade de riscos.

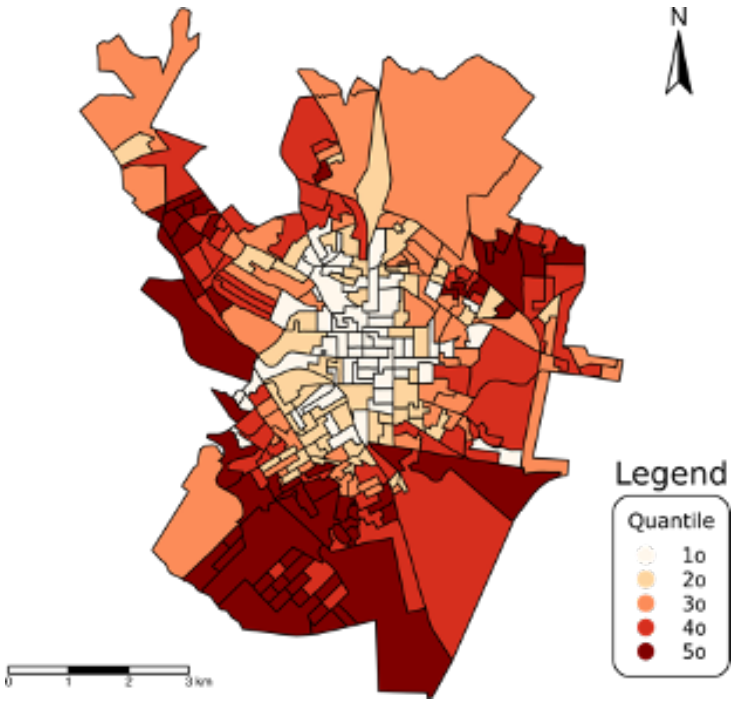
O Índice Paulista de Vulnerabilidade Social (IPVS) é um índice concebido pela Fundação Sistema Estadual de Análise de Dados (SEADE) do estado de SP, e implementado a partir do conjunto de informações existentes no banco de dados do universo do Censo Demográfico 2010, do IBGE que consiste nas informações socioeconômicas e demográficas investigadas pelo Censo, e agregadas no nível de setores censitários (SEADE, 2010).

Segundo a metodologia adotada, os setores censitários com pelo menos 50 domicílios particulares permanentes foram classificados em um dos seis grupos: Grupo 1 – baixíssima vulnerabilidade; Grupo 2 – vulnerabilidade muito baixa; Grupo 3 – vulnerabilidade baixa; Grupo 4 – vulnerabilidade média; Grupo 5 – vulnerabilidade alta; e Grupo 6 – vulnerabilidade muito alta.

O Grupo 6 (vulnerabilidade muito alta) engloba apenas setores censitários classificados no Censo Demográfico como aglomerados subnormais com concentração de população jovem e de baixa renda.

O IPVS consiste em uma tipologia de situações de exposição à vulnerabilidade, agregando aos indicadores de renda, outros referentes ao ciclo de vida familiar e escolaridade, no espaço intraurbano. Com efeito, entre as questões investigadas pelo Censo Demográfico 2010 em seu

Figura 48 – Representação coroplética da distribuição da média de moradores por domicílio.



Fonte: Elaborada pelo autor.

questionário básico, além das variáveis socioeconômicas (renda e condição de alfabetização), elegeram-se às relacionadas ao ciclo de vida familiar (presença de crianças menores, idade e gênero do chefe de família). Assim, os componentes das duas dimensões do IPVS (demográfica e social) estão descritos na Figura 51.

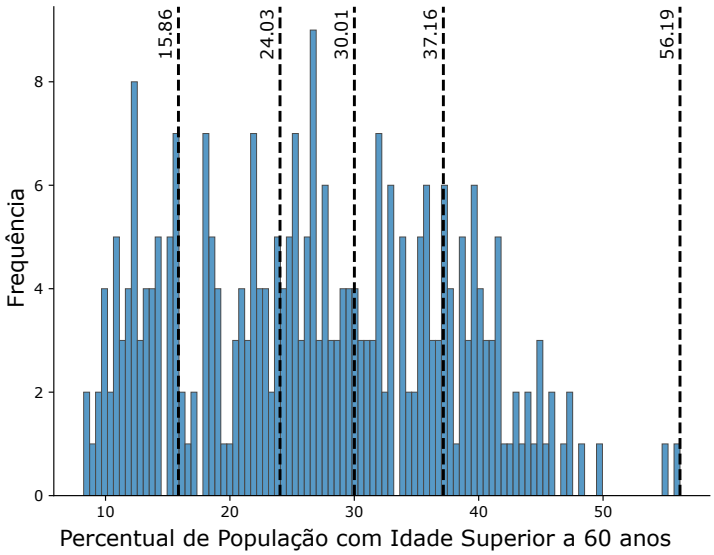
Os *scores* para cada setor receberam valores de 1 a 6 variando da baixíssima vulnerabilidade até a alta ou muito alta vulnerabilidade. Na Figura 52 pode-se visualizar a distribuição do IPVS nos setores.

Agravos de saúde

A prévia existência de casos de um agravo de saúde, em uma determinada região, coloca-a sob uma condição de risco ainda mais grave. Quanto maior o número desses casos, pior a condição da região.

Nesse estudo de caso, foram considerados os seguintes agravos de saúde: (i) casos de dengue notificados e confirmados pelo Sistema de Informação de Agravos de Notificação (SINAN-Dengue); (ii) casos confirmados de COVID-19 no Sistemas de Informação de Vigilância Epidemiológica da Gripe (SIVEP-gripe); e (iii) casos de tuberculose registrados no Sistema de Controle de Pacientes com Tuberculose (TBWeb) do Estado de São Paulo. O período de análise foi de 1º de janeiro de 2020 a 31 de dezembro de 2020, e os dados foram disponibilizados pela VIGEP-SC.

Figura 49 – Histograma dos dados de percentual de população com idade superior a 60 anos.



Fonte: Elaborada pelo autor.

A Figura 53 mostra, para cada semana epidemiológica, o número de notificações de casos de dengue, COVID-19 e tuberculose, respectivamente.

Todos os dados dos agravos de saúde foram inseridos no MultiMapas através do módulo de geolocalização, seguindo o *workflow* de ingestão do módulo (Figura 31). Após a ingestão dos dados dos múltiplos agravos de saúde, foi realizado a contagem de cada agravo de saúde, nos setores censitários, através do módulo de agregação do MultiMapas.

Na Figura 54, pode-se observar a distribuição espacial dos casos confirmados de dengue em 2020. Na Figura 55, apresenta-se a distribuição espacial dos casos de COVID-19 em 2020, e na Figura 56, temos a distribuição espacial de casos de tuberculose em 2020.

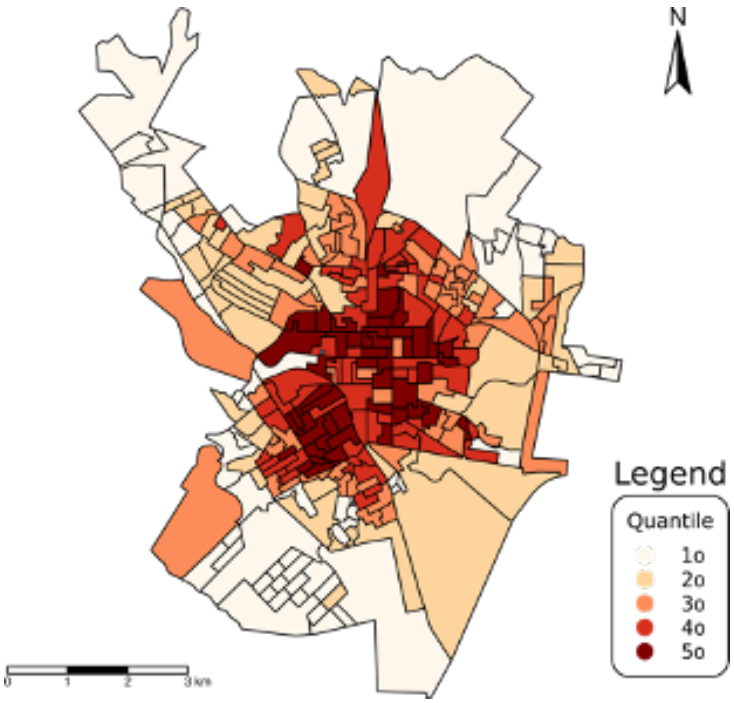
Os *scores* de cada setor é a própria contagem de casos.

5.2.4 Avaliação experimental

Os experimentos dessa seção visam analisar o desempenho do MultiMapas. Os experimentos foram conduzidos em um computador com processador Apple M1 Max com 32 GB de RAM e 2 TB de armazenamento em disco e sistema operacional macOS 13.4 Ventura.

Como visto até aqui, compreender e mapear os riscos, problemas e dificuldades no enfrentamento de um agravo à saúde é complexo e envolve diversas variáveis com características diversas e muitas vezes de fontes difusas. Escolher alternativas e estabelecer um modelo racional para combinar os dados é necessário. Nesse contexto, o geoprocessamento pode oferecer ferramentas de apoio à tomada de decisão que auxiliem o especialista nessa tarefa.

Figura 50 – Representação coroplética da distribuição do percentual de população com idade superior a 60 anos.



Fonte: Elaborada pelo autor.

O conceito fundamental dos vários modelos de tomada de decisão é o de racionalidade. Segundo esse princípio, indivíduos e organizações seguem o comportamento de escolher entre alternativas com base em critérios objetivos de julgamento, cujo fundamento será o de satisfazer um nível de desejo pré-estabelecido (CLEMEN; REILLY, 2013).

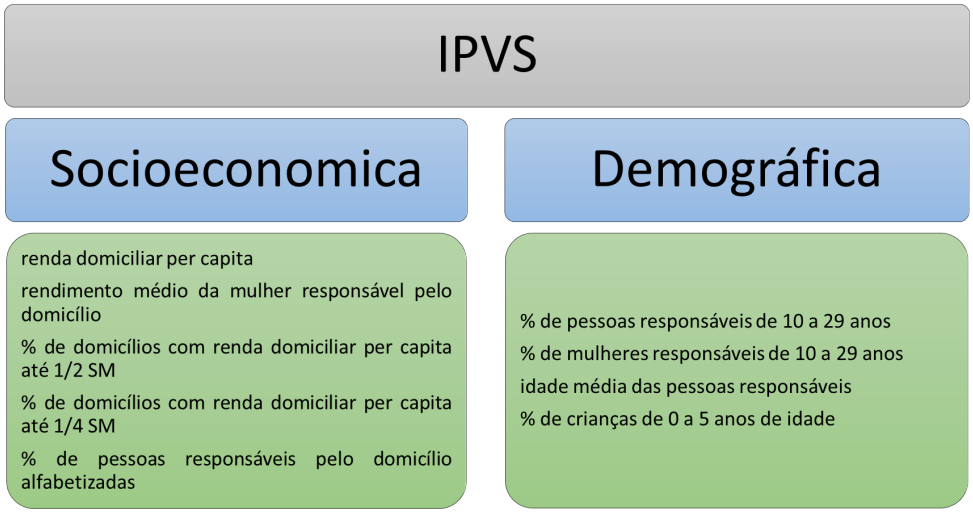
Para determinar a contribuição relativa de cada variável (camadas temáticas), como mencionado na Seção 5.2.2, primeiro decidiu-se classificá-las em três grupos de acordo com aspectos demográficos, sociais e epidemiológicos. Na Figura 57 é apresentada a classificação das variáveis segundo esses aspectos.

Uma vez definido o conjunto de camadas temáticas e como esse pode ser agrupado, foi utilizado o módulo *decision making* do MultiMapas, considerando a definição do mapa global final conforme a Equação 5.1:

$$\begin{aligned} \mu_i = {} & g_1 \times \begin{pmatrix} d_1 \times DemoDens_i + d_2 \times MorPDom_i + \\ d_3 \times PercPop60_i \end{pmatrix} + \\ & g_2 \times \left(s_1 \times IPVS_i \right) + \\ & g_3 \times \begin{pmatrix} e_1 \times HUCount_i + e_2 \times dengueCount_i + \\ e_3 \times covidCount_i + e_4 \times tbCount_i \end{pmatrix} \end{aligned} \tag{5.1}$$

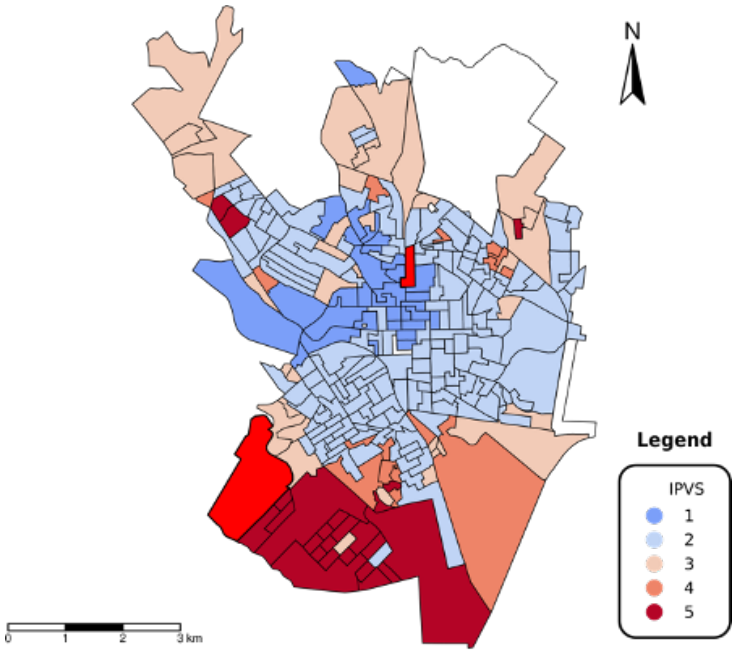
no qual g_i, com $i \in [1,2,3]$, representa os pesos de cada grupo de cada camadas temáticas, d_j,

Figura 51 – Quadro-resumo das variáveis componentes do IPVS, segundo suas dimensões (SEADE, 2010).



Fonte: Elaborada pelo autor.

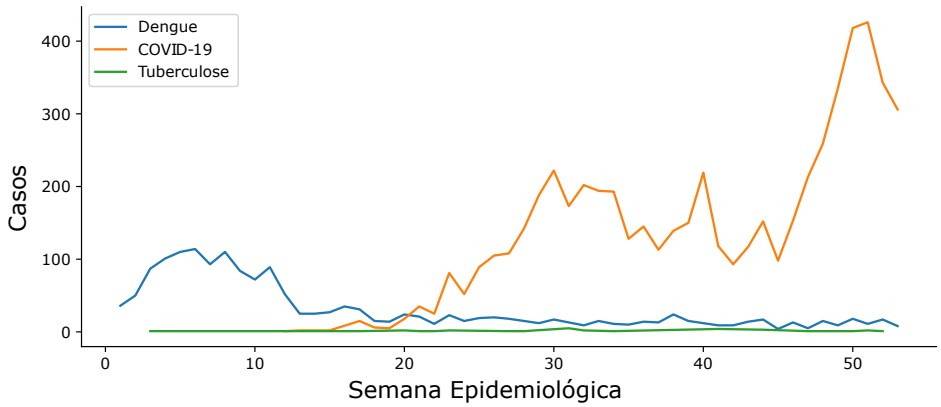
Figura 52 – Representação coroplética da distribuição do IPVS.



Fonte: Elaborada pelo autor.

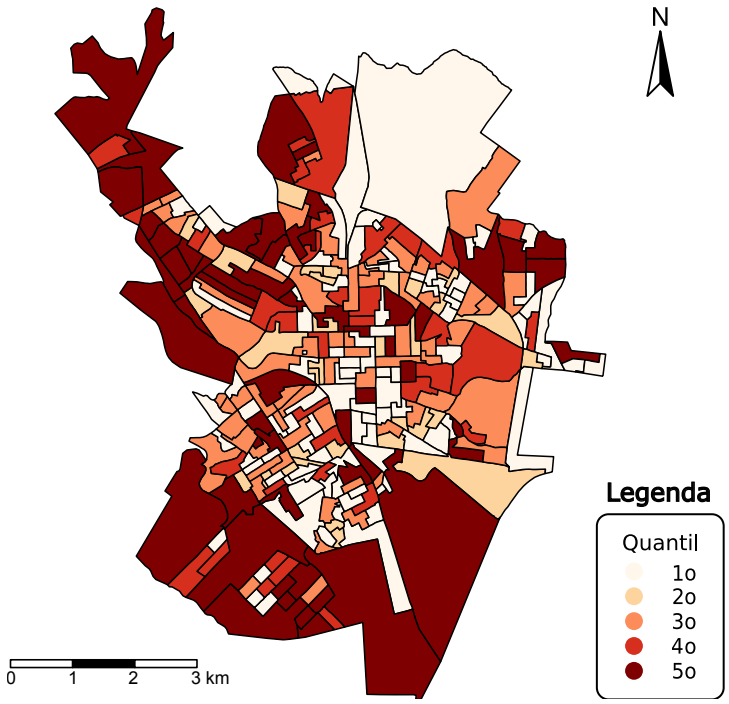
com $j \in [1,2,3]$, os pesos das camadas temáticas com informações demográficas, s_1 o peso da camada temáticas com aspectos sociais e e_k, com $k \in [1,2,3,4]$, os pesos das camadas temáticas com dados epidemiológicos.

Figura 53 – Notificações de Casos de Dengue, COVID-19 e Tuberculose por Semana Epidemiológica de 2020.



Fonte: Elaborada pelo autor.

Figura 54 – Representação coroplética da distribuição espacial dos casos notificados de Dengue em 2020.



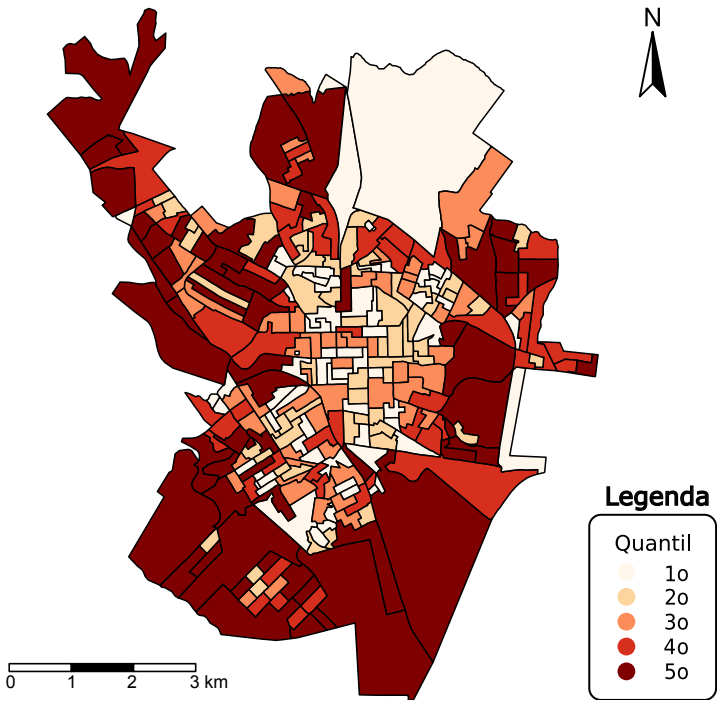
Fonte: Elaborada pelo autor.

5.2.5 Método AHP

A tarefa de entender e mapear os riscos, problemas e dificuldades de enfrentamento de múltiplos agravos de saúde é complexo e envolve diferentes variáveis de características diversas e muitas vezes de fontes difusas.

Inicialmente, foi utilizado o método AHP para definição de uma solução a partir da visão do grupo de especialista consultado. A cada grupo foi solicitado que fizesse a comparação

Figura 55 – Representação coroplética da distribuição espacial dos casos notificados de COVID-19 em 2020.



Fonte: Elaborada pelo autor.

pareada entre os grupos de cada camada temática a partir da escala fundamental de Saaty (1978) para a definição de seus pesos através do MultiMapas, considerando como resultado a comparação que mais se repetia entre os especialistas.

A Tabela 10 apresenta a matriz de comparação pareada e os respectivos pesos de cada grupo. A Tabela 11 apresenta a matriz de comparação pareada e os respectivos pesos para cada variável.

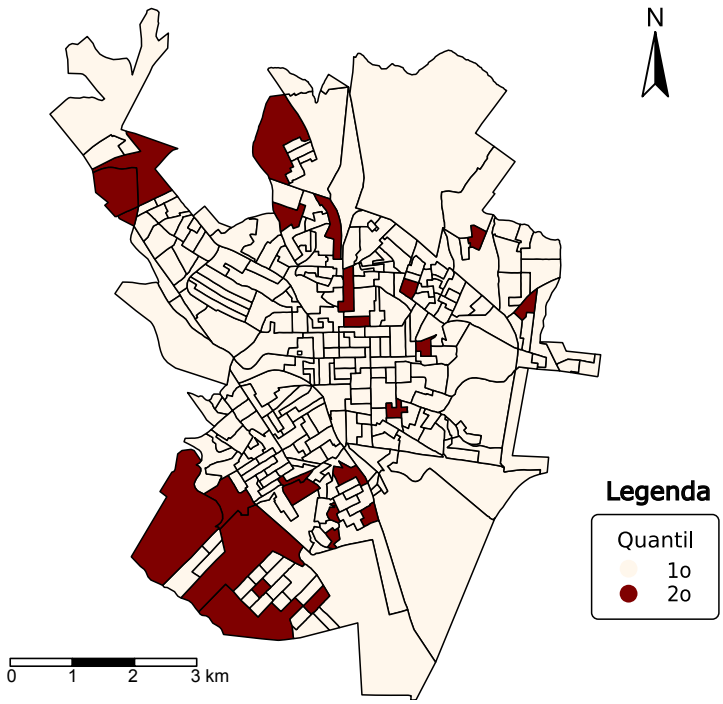
Tabela 10 – Matriz de comparação pareada para grupos.

Grupo	**Matriz Pareada**	**A**	**B**	**C**	**Pesos**
Demográficos	A	1.00	3.00	0.50	0.31
Social	B	0.33	1.00	0.20	0.11
Epidemiológicos	C	2.00	5.00	1.00	0.58

Fonte: Dados da pesquisa.

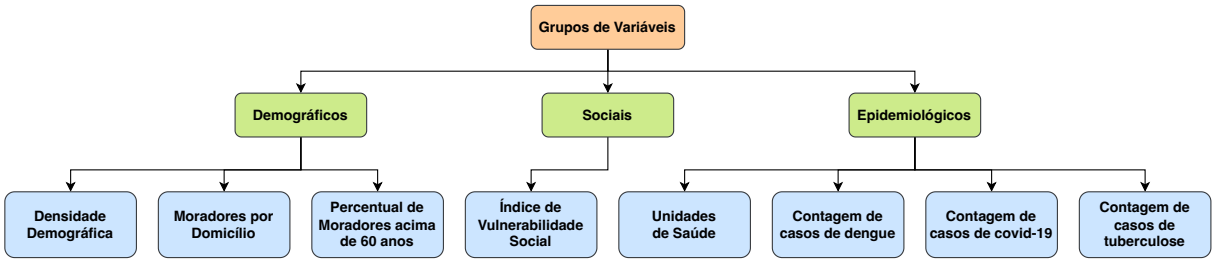
Dada as matrizes de comparação pareada dos grupos e variáveis, foi calculada a Razão de Consistência (do inglês *Consistency Ratio* - CR) para validar as comparações pareadas entre grupos e variáveis, conforme proposto por Saaty (1977). A CR para grupos de variáveis e variáveis demográficas foi de 0,005 ($CR_{grupo} = CR_{var\_dem} 0,005$), enquanto para variáveis epidemiológicas foi de 0,21 ($CR_{var\_epi} = 0,21$). Assim, o valor de CR ≤ 0.01 (SAATY; VERGAS, 1992) mostra que os valores de prioridade relativa são consistentes.

Figura 56 – Representação coroplética da distribuição espacial dos casos notificados de tuberculose em 2020.



Fonte: Elaborada pelo autor.

Figura 57 – Classificação das camadas temáticas em grupos segundo seus aspectos.



Fonte: Elaborada pelo autor.

A partir desses resultados, tem-se que o grupo de variáveis epidemiológicas aparece com maior nível de influência para a determinação de áreas críticas. Relativo às variáveis demográficas, a variável que representa a percentagem da população com mais de 60 anos parece ter o maior nível de influência relativamente às restantes variáveis do mesmo grupo. Isso ocorreu, segundo o grupo de especialista consultado, devido às pessoas nessa faixa etária terem dificuldades de mobilidade mais significativas e por serem mais suscetíveis a doenças pulmonares.

Para as variáveis epidemiológicas, a variável com influência mais significativa é a variável que representa o número de casos de dengue. Essa importância se dá, segundo o grupo de especialistas consultado, devido o ser um vetor (mosquito *Aedes aegypti*) o transmissor desse agravo e alguns sintomas se confundem com os de outras doenças (ZHANG *et al.*, 2023). Na Figura 58, podemos visualizar a composição final dos pesos para o conjunto completo de

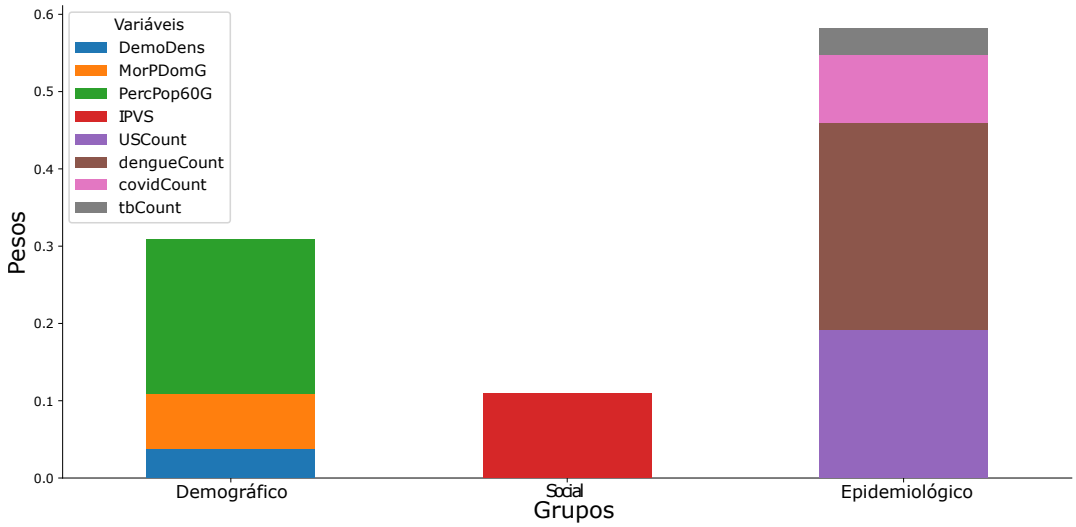
Tabela 11 – Matriz de comparação pareada para as variáveis.

Grupo	Variável	#	A	B	C	D	Pesos
Demográficos	DemoDens	A	1.00	0.50	0.20	-	0.12
	MorPDom	B	2.00	1.00	0.33	-	0.23
	PercPop60	C	5.00	3.00	1.00	-	0.65
Social	IPVS	A	1.00	-	-	-	1.00
Epidemiológicos	HUCount	A	1.00	2.00	2.00	2.00	0.33
	dengueCount	B	0.50	1.00	6.00	6.00	0.46
	covidCount	C	0.50	0.17	1.00	6.00	0.15
	tbCount	D	0.50	0.17	0.17	1.00	0.06

Fonte: Dados da pesquisa.

variáveis.

Figura 58 – Composição final dos pesos do método AHP.



Fonte: Elaborada pelo autor.

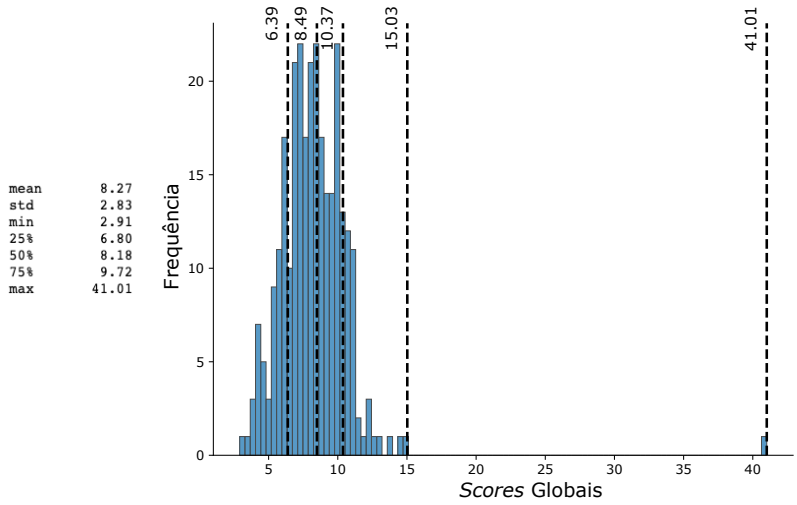
Com a definição do peso de cada variável no modelo, calcula-se um *score* global cruzando as diferentes camadas que representam cada variável ponderada pelos respectivos pesos usando uma combinação linear ponderada (WLC), expressa pela Equação 5.2, na qual a camada temática de adequação final é derivada multiplicando cada camada temática por seu peso relativo seguido pela soma dos resultados.

$$\begin{aligned} \mu_i = {} & 0.31 \times \begin{pmatrix} 0.12 \times DemoDens_i + 0.23 \times MorPDom_i + \\ 0.65 \times PercPop60_i \end{pmatrix} + \\ & 0.11 \times \left(1.00 \times IPVS_i\right) + \\ & 0.58 \times \begin{pmatrix} 0.33 \times HUCount_i + 0.46 \times dengueCount_i + \\ 0.15 \times covidCount_i + 0.06 \times tbCount_i \end{pmatrix} \end{aligned} \tag{5.2}$$

A partir dos *scores* globais calculados, foi realizada a primeira análise para verificar suas estatísticas descritivas. Estas são mostradas na Figura 59.

Os valores globais *scores* foram então divididos em cinco classes de acordo com seus valores naturais de quebra. Nesse caso, o método de classificação adotado tem a característica de agrupar valores semelhantes e maximizar as diferenças entre as classes com limites estabelecidos, nos quais existem diferenças consideráveis entre os valores dos dados. Assim, esse método representa o escalonamento natural das séries de dados, agrupando-as de acordo com (MATSUMOTO; CATÃO; GUIMARÃES, 2017) similaridade. A Figura 60 apresenta a classificação dos setores censitários em relação aos *scores* globais do modelo.

Figura 59 – Histograma com as estatísticas descritivas do *Score* Global gerada pelo método AHP.



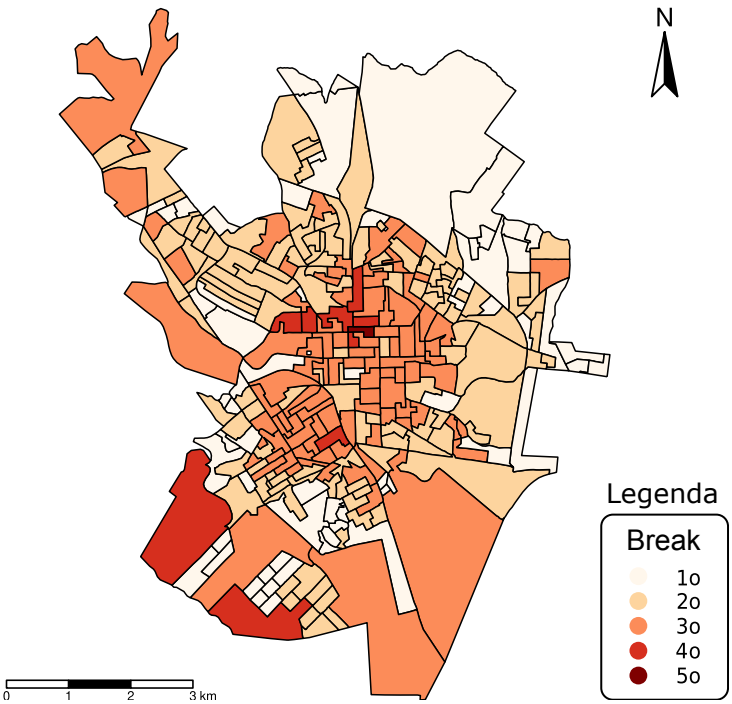
Fonte: Elaborada pelo autor.

Para certificar-se que a modelagem pelo método AHP representa um fenômeno do ponto de vista espacial, e assim validarmos o mapeamento em questão, um dos aspectos necessários que temos que responder é se o evento, em estudo, e os fatores relacionados a ele possuem distribuição espacialmente condicionada. Nesse caso, tem-se que usar a estatística espacial, em particular, o estudo da dependência espacial para demonstrar como os valores estão correlacionados com o espaço, ou seja, se e como dependem de valores da mesma variável nas regiões vizinhas.

Para essa verificação foi calculado o índice global *I* de Moran, dada pela Equação 2.5. Os valores de *scores* globais de cada setor foram usados para calcular o Moran I. Na Figura 61 pode-se verificar o resultado do teste. Dado o *z-score* de 10.009 e com índice *I* de Moran igual a 0,31, indicando que há uma probabilidade inferior a 1% de que esse padrão de agrupamento possa ser um resultado aleatório. Assim, a distribuição espacial de valores altos e/ou baixos no conjunto de dados é mais espacialmente agrupada do que seria esperado se os processos espaciais subjacentes fossem aleatórios.

Indicadores globais como o Moran I fornecem um único valor como medida da associação

Figura 60 – Representação coroplética da distribuição da classificação dos setores censitários em relação a esses *scores* utilizando o método de escalonamento natural.



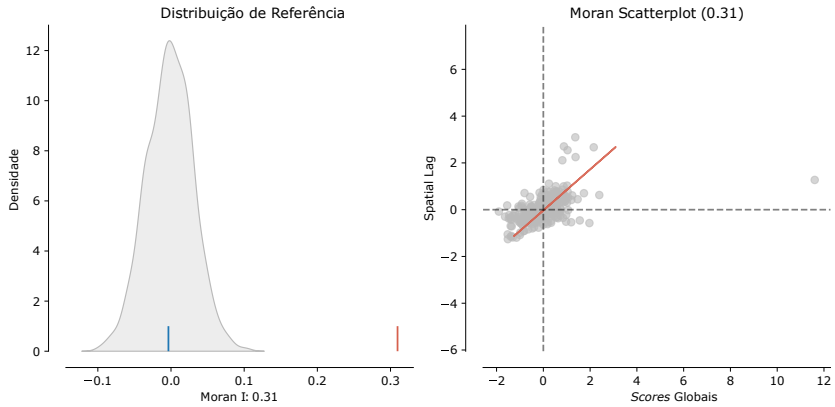
Fonte: Elaborada pelo autor.

espacial para todo o conjunto de dados, o que é útil na caracterização da região na totalidade. Assim pode-se rejeitar a hipótese nula, mas não se tem ainda como analisar significativamente o padrão dos agrupamentos (CÂMARA *et al.*, 2004).

Utilizando os centroides dos setores censitários como ponto de partida, pode-se variar a distância de vizinhança a outros centroides para calcular o Moran I e assim analisar como a dependência espacial está variando em função dessa distância. Para isso, analisamos par a par as distâncias entre os centroides e verificamos que a distância mínima encontrada foi da ordem de 130 metros e a máxima da ordem de 1600 metros com uma média de distância da ordem de 300 metros. Desta forma, calculou-se o Moram I para distâncias a cada 150 metros partindo de 300 metros, em 30 intervalos. Na Figura 62, apresentamos o gráfico da variação do *z-score* em função da distância. Os picos refletem as distâncias em que os processos espaciais que promovem o agrupamento são mais pronunciados. A cor de cada ponto, no gráfico, corresponde à significância estatística dos valores do *z-score*.

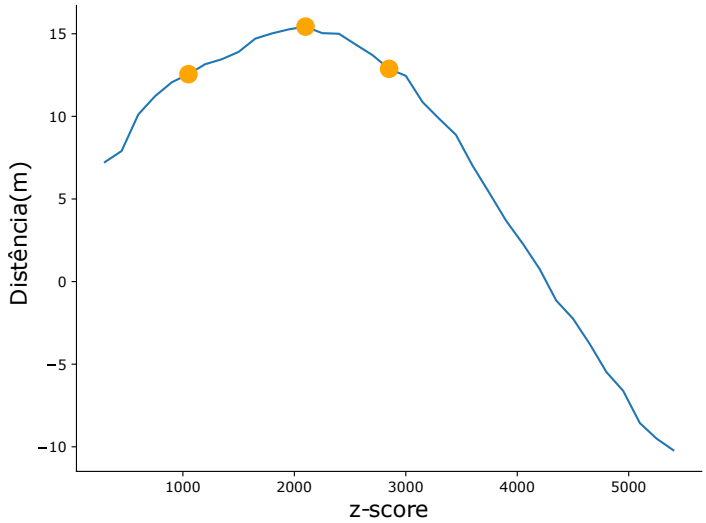
A partir do mapa temático modelado pelo método AHP (Figura 60), é possível observar que temos duas áreas críticas, a região central e sul do município. A criticidade da região central se deu pelo fato de que é a região do município com a maior concentração de pessoas acima de 60 anos, ver Figura 50, e a variável considerada com maior peso dentre as variáveis demográficas. Já em relação à região sul, observa-se, nela, uma grande quantidade de setores censitários com IPVS 5, ver Figura 52, uma região com alta vulnerabilidade social, além de ser a região do

Figura 61 – Resultado do teste I de Moran dos *scores* globais gerados pelo método AHP.



Fonte: Elaborada pelo autor.

Figura 62 – Autocorrelação especial em função da distância desses *scores*.



Fonte: Elaborada pelo autor.

município com a maior concentração dos agravos considerados, ver as Figuras 54, 55 e 56, evidenciando a capacidade do modelo em captar essas nuanças.

Apesar da solução gerada pelo método AHP representar um fenômeno do ponto de vista espacial e o método AHP ser numericamente método MCDM mais utilizado para integrar GIS e MCDA (MALCZEWSKI; RINNER, 2015), o seu uso possui algumas limitações, como: (i) Segundo Odu (2019), o método AHP depende da visão de um especialista para poder utilizar os critérios de julgamento, a escala fundamental de Saaty (1977), para definir o grau de importância de cada critério ou subcritério utilizado. (ii) um grande número de comparações pareadas quando se tem muitos critérios. Uma das grandes desvantagens do método AHP consiste no número de comparações por pares que podem ser feitas. Isso ocorre porque é necessário comparar os critérios individuais e, em seguida, comparar as alternativas em relação a um critério específico. O método

AHP gera $\frac{n^2-n}{2}$ comparações a partir de n critérios e como cada critério pode assumir até duas posições na matriz de comparações e a escala fundamental possui nove opções de julgamento, o número de matriz de julgamento pareada que podem ser geradas é igual a $2^{\frac{n^2-n}{2}} \times 9^{\frac{n^2-n}{2}}$. Se o problema de tomada de decisão tiver um grande número de critérios, bem como alternativas, o processo de tomada de decisão pode ser demorado e, com o tempo, a inconsistência das matrizes pode começar a aumentar devido à perda de atenção e falta de concentração do assunto (MUNIER; HONTORIA *et al.*, 2021). (iii) O método AHP aplicado, dessa forma, não otimiza a dependência espacial.

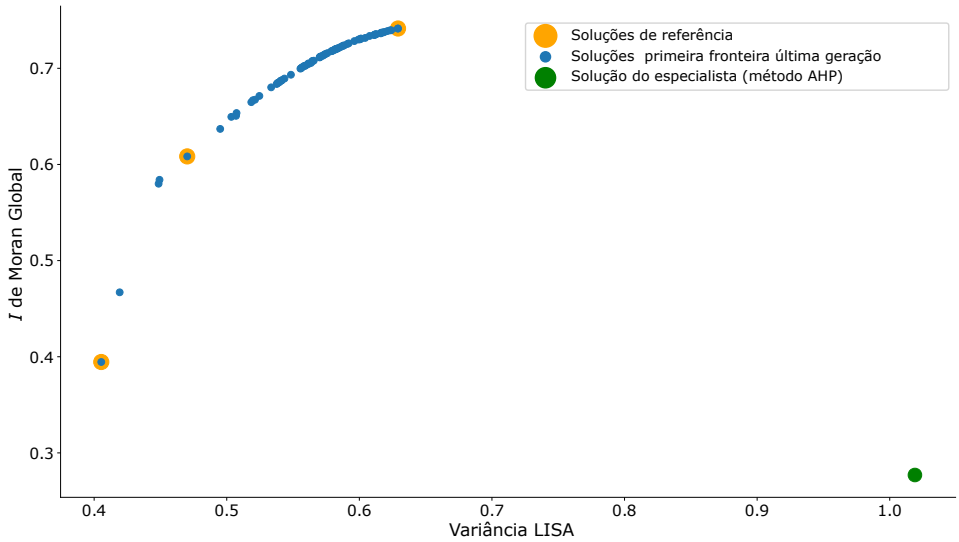
Nesse contexto foi utilizado GIS-moGA, um algoritmo genético multiobjetivo baseado no NSGA-II (DEB *et al.*, 2002), que visa mitigar as limitações do método AHP para a construção de uma representação cartográfica derivada da composição de múltiplos mapas temáticos.

5.2.6 GIS-moGA

O GIS-moGA foi executado 30 vezes com os seguintes parâmetros: alfa = uniforme aleatório $\in$ [0, 1/2], taxa de mutação = 0,01, população inicial = 100, número máximo de gerações = 300. Esses valores foram definidos empiricamente. A Equação 5.1 foi utilizada para produzir *trade-off* de Pareto entre a maximização do índice *I* de Moran (dependência espacial) e a minimização do índice local de Moran LISA (heterogeneidade espacial).

A Figura 63 mostra os resultados da primeira fronteira encontrada pelo GIS-moGA na última geração do experimento. Analisamos o número de fronteiras retornadas, durante o processo de otimização, e, apesar de termos encontrado quatro fronteiras, essas se sobrepuseram à primeira. Portanto, não houve melhora na fronteira de Pareto na última geração.

Figura 63 – Soluções da primeira fronteira encontrada pelo GIS-moGA na última geração.



Fonte: Elaborada pelo autor.

Ao resolver problemas de otimização multiobjetivo, estamos potencializando múltiplos objetivos conflitantes simultaneamente. Devido à natureza conflitante dos objetivos, não podemos melhorar o valor de nenhum objetivo sem prejudicar pelo menos um dos outros objetivos (ESKELINEN; MIETTINEN, 2012). Um trade-off representa a desistência de um dos objetivos, o que permite a melhoria de outro objetivo. No entanto, precisamos apoiar o tomador de decisão na escolha da solução ideal para o problema. Para isso, apresenta-se um *trade-off* de três soluções, chamadas de soluções de referência, a partir do conjunto de soluções ótimas encontradas pelo GIS-moGA na primeira fronteira de Pareto da última geração que considera o relaxamento de um dos objetivos em relação ao outro e uma solução intermediária. Na Figura 63, destacam-se essas três soluções de referência da primeira fronteira encontrada pelo GIS-moGa na última geração, sendo uma considerando apenas a maximização do I de Moran Global, outra considerando a minimização da variância do LISA, e outra que consiste na solução intermediária entre a maximização do I de Moran Global e a minimização da variância do LISA. É possível também observar, na Figura 63, que as soluções da primeira fronteira do GIS-moGA na última geração do ponto de vista da dependência e heterogeneidade espacial são melhores do que a solução encontrada pelo método AHP a partir da Equação 5.2.

Para avaliar o desempenho do GIS-moGA, foi utilizado o indicador de hipervolume (HV), uma métrica amplamente utilizada na avaliação de algoritmos multiobjetivos (DEB, 2011). No HV, é calculado o volume da região coberta entre os pontos das soluções na fronteira de Pareto P e um ponto de referência W. Para cada solução $i \in P$, é construído um hipercubo, v_i, com referência a um ponto W. Esse ponto de referência pode ser encontrado construindo um vetor com os piores valores da função objetivo. A união de todos os hipercubos encontrados é o resultado da métrica. Quanto maior o valor de HV melhor. Um alto valor de HV indica que houve um alto *spread* entre as soluções P e indica que também houve uma melhor convergência. O HV é calculado conforme a Equação 5.3.

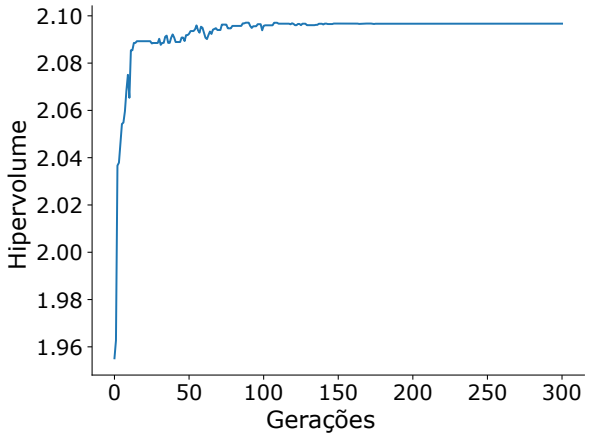
$$HV = \sum_{i \in P} v_i \tag{5.3}$$

A Figura 64 mostra o hipervolume gerado pelo conjunto de soluções não dominadas em cada geração em nosso GIS-moGA.

Com base nas soluções de referência, foi realizada a determinação do algoritmo que melhor descreve número de classes e os limites de cada classe da distribuição do *score global* para construção da representação coroplética. Essa escolha foi realizada a partir do cálculo do ADCM dos *scores* globais, como pode ser observado nas Figuras 65, 66 e 67 de cada solução de referência. Como pode ser observado, o menor ADCM, em cada solução de referência encontrado, foi o algoritmo *Jenks Caspall*.

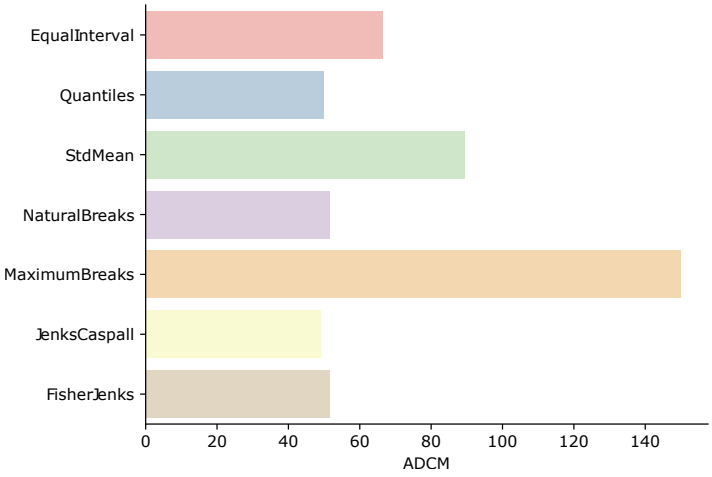
Uma vez determinado o algoritmo que melhor descreve o *score global* de cada solução de referência, pode-se analisar a distribuição desses valores. Na Tabela 12 temos das estatísticas

Figura 64 – Hipervolume gerado pelo conjunto de soluções não dominadas em cada geração do GISmoGA.



Fonte: Elaborada pelo autor.

Figura 65 – ADCM do *Score Global* das soluções de referência Inferior.



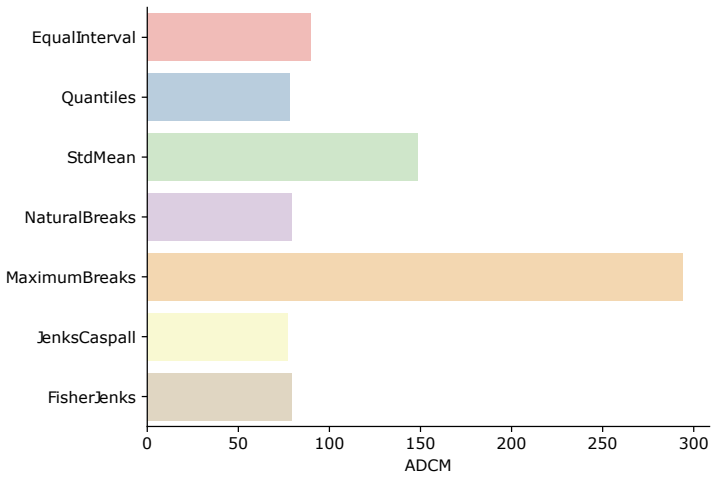
Fonte: Elaborada pelo autor.

descritivas dos *scores globais* de cada solução de referência.

As Figuras 68, 69 e 70 mostram os histogramas com as frequências dos *scores* globais de cada solução de referência nos setores censitários. Já as Figuras 71, 72 e 73, apresentam as representações coropléticas dos *scores* globais de cada solução de referência por setor censitário, segundo o algoritmo de *Jenks Caspall*.

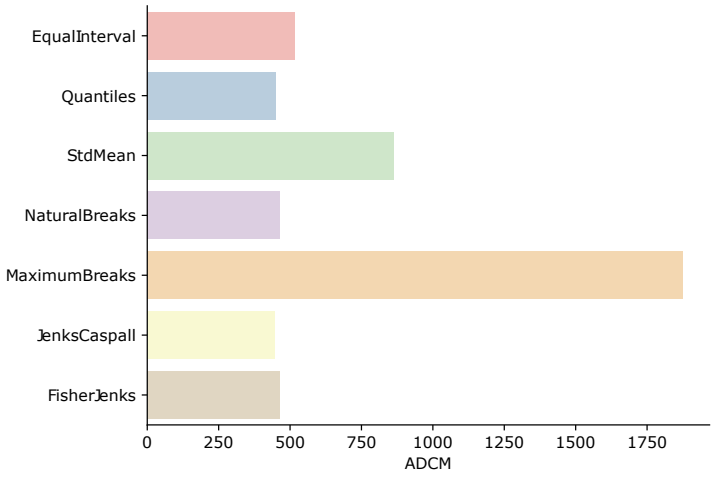
Nas Figuras 74, 75 e 76, temos o Moran *Scatterplot* que mostra o *lag* espacial do *score global* das soluções de referência. É possível observar que a solução de maximização do I de Moran Global apresenta um número mais significativo de regiões com associações lineares espaciais alto-alto (HH) e baixo-baixo (LL), com diminuição desses tipos de associações para a

Figura 66 – ADCM do *Score Global* das soluções de referência Intermediária.



Fonte: Elaborada pelo autor.

Figura 67 – ADCM do *Score Global* das soluções de referência Superior.



Fonte: Elaborada pelo autor.

solução intermediária e minimização da variância LISA. Vale ressaltar que a associação linear espacial do tipo HH significa que as regiões apresentam valores elevados da variável de interesse, ou seja, regiões vizinhas acima da média que também apresentam valores elevados (ALMEIDA, 2012; LOPES *et al.*, 2022).

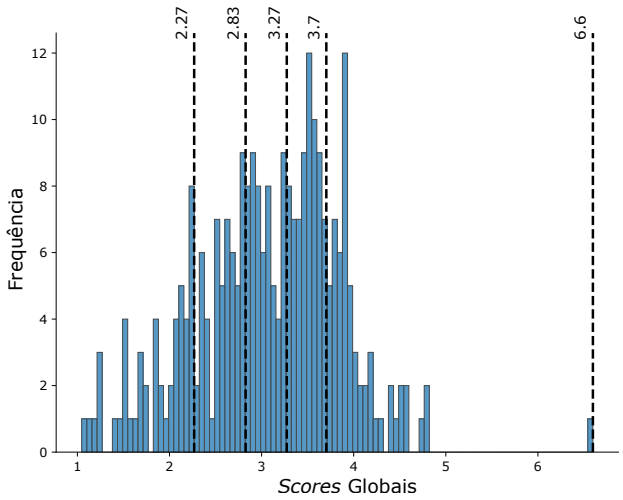
As Figuras 77, 78 e 79, mostram os mapas com os *clusters* LISA das soluções de referência. É possível observar que a solução que maximiza o I de Moran Global apresenta na região central a maior concentração de regiões com associação espacial linear do tipo HH. Vale ressaltar que a região central possui a maior concentração de pessoas com mais de 60 anos (Figura 50), grupo mais suscetível a contrair lesões mais graves, além de apresentar, em geral,

Tabela 12 – Estatísticas Descritivas das Soluções de Referência.

	Score Global Solução de Referência Inferior	**Score Global Solução de Referência Intermediária**	**Score Global Solução de Referência Superior**
Média	3,07	5,15	22,63
Desvio Padrão	0,79	1,36	8,37
Mínimo	1,05	1,98	7,15
25%	2,56	4,14	15,60
50%	3,13	5,17	22,46
75%	3,62	6,24	49,39
Máximo	6,60	8,48	45,65

Fonte: Dados da pesquisa.

Figura 68 – Histograma do *Score Global* das soluções de referência Inferior.



Fonte: Elaborada pelo autor.

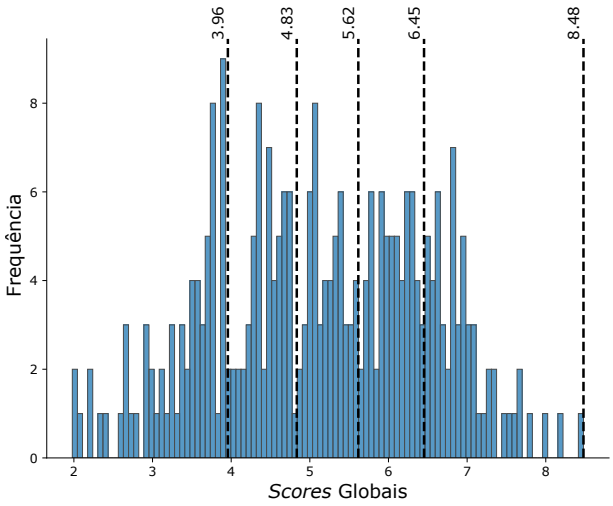
grande dificuldade de locomoção, sendo, portanto, uma região crítica do ponto de vista dos agravos de saúde.

Para verificarmos o grau de importância das variáveis, camadas temáticas, em cada solução de referência que foram considerados pelo GIS-moGA, é necessário primeiro verificar os valores do *I* de Moran Global e a variância do índice de Moran local, LISA, para cada uma das variáveis. A Figura 80 apresenta o valor do índice *I* de Moran Global e a variância do índice de Moran local, LISA, das oito variáveis que foram consideradas pelo GIS-moGA no processo de otimização.

Na Tabela 13, temos os pesos, o grau de importância, dados pelo GIS-moGA para cada variáveis de referência.

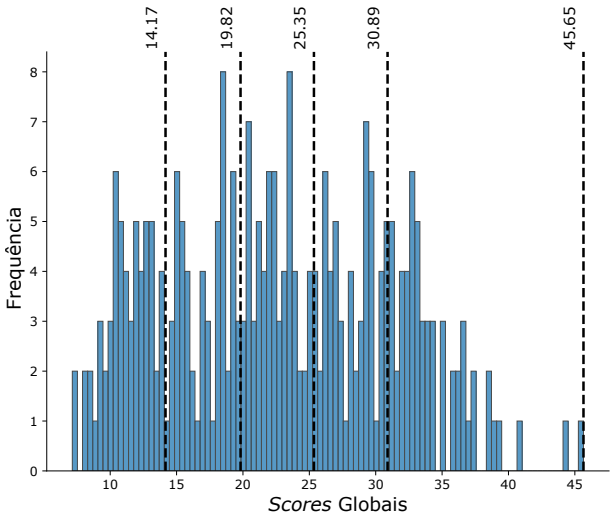
Analisando os dados da Tabela 13, temos que o GIS-moGA desconsiderou as variáveis *HCCScore*, *DengueCount* e *CovidCoun*, nas três soluções de referência, e ao observar a Figura 80

Figura 69 – Histograma do *Score Global* das soluções de referência Intermediária.



Fonte: Elaborada pelo autor.

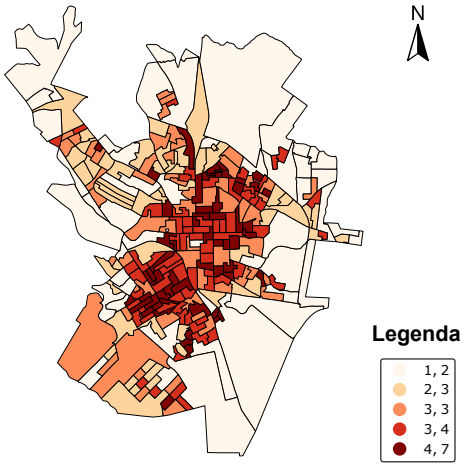
Figura 70 – Histograma do *Score Global* das soluções de referência Superior.



Fonte: Elaborada pelo autor.

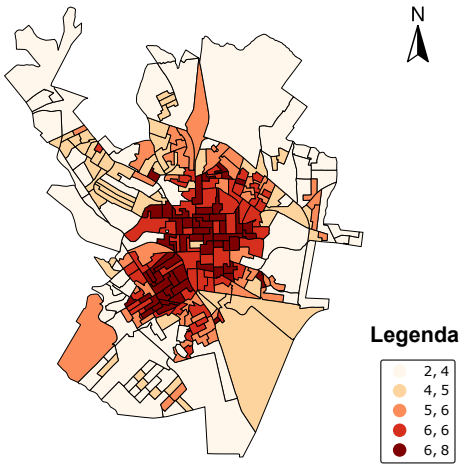
a variável *DengueCount* é a que possui o menor I de Moran Global e o menor LISA entre as demais variáveis, o que faz com que possamos concluir que essa variável não influencia significamente na dependência e heterogeneidade espacial global, já a variável *HCCScore* tenha um I de Moran Global maior entre as variáveis epidemiológicas e a segunda variância do índice LISA maior entre esse mesmo grupo, ela não influencia significamente na dependência e heterogeneidade espacial global. Já as variáveis *DensDemoG*, *PercPop*60*G* e *TbCount* não foram desconsideradas nas três soluções de referência, o que faz com que possamos concluir que essas variáveis contribuem, de uma certa forma, para a dependência e heterogeneidade espacial global.

Figura 71 – Representação coroplético dos *Scores* Globais segundo o algoritmo *Jenks Caspall* das soluções de referência Inferior.



Fonte: Elaborada pelo autor.

Figura 72 – Representação coroplético dos *Scores* Globais segundo o algoritmo *Jenks Caspall* das soluções de referência Intermediária.



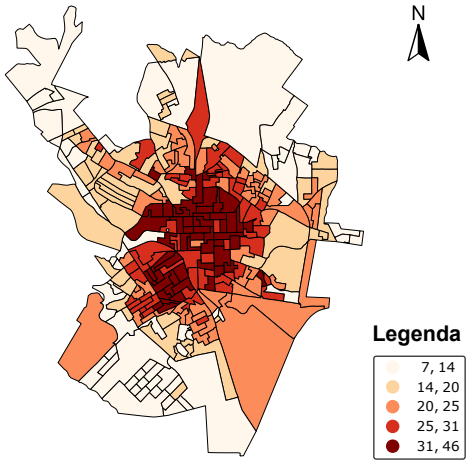
Fonte: Elaborada pelo autor.

Na Figura 81, temos um *heatmap* dos pesos encontrados pelo GIS-moGA das variáveis na primeira fronteira de Pareto da última geração. É possível observar que as variáveis *HCCScore*, *DengueCount* e *CovidCoun* não foram consideradas em nenhuma das soluções da primeira fronteira da última geração do GIS-moGA, ou seja, essas variáveis não contribuem na dependência e heterogeneidade espacial global.

5.3 Estudo de caso 2: Seguros agrícolas para tomate

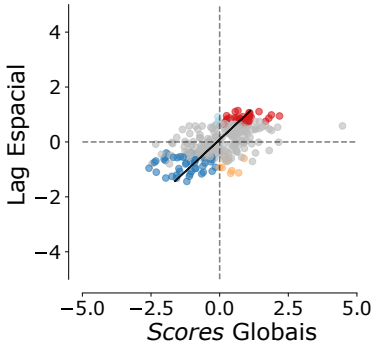
Os seguros agrícolas são uma parte vital das cadeias de abastecimento agroalimentar. Proporcionam segurança ao agricultor, permitindo investimentos na produção agrícola. Eles

Figura 73 – Representação coroplético dos *Scores* Globais segundo o algoritmo *Jenks Caspall* das soluções de referência Superior.



Fonte: Elaborada pelo autor.

Figura 74 – Moran *Scartterplot* dos *Scores* Globais das soluções de referência Inferior.



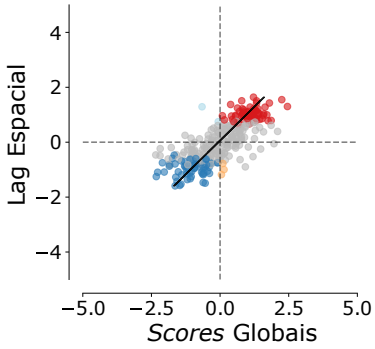
Fonte: Elaborada pelo autor.

também estão sendo propostos como medida para reduzir os impactos da ocorrência de eventos climáticos extremos como secas, enchentes e perigos em áreas agrícolas (IPCC, 2021; KIM; IIZUMI; NISHIMORI, 2019). Conforme explorado nos trabalhos de (IPCC, 2021; SEIFERT-DÄHNN, 2018), eventos climáticos extremos podem ter impactos significativos na quantidade, qualidade e distribuição de produtos, impactando diretamente na sustentabilidade das diferentes cadeias agroalimentares. Uma melhor compreensão das diferentes políticas e tendências é crucial para aumentar a adoção desta importante ferramenta de transferência de risco, melhorando a resiliência das cadeias de abastecimento agroalimentar.

O presente estudo de caso foi solicitado por um grupo de pesquisadores multidisciplinares que usam inteligência artificial e modelos estatísticos para melhor abordar problemas relacionados à avaliação da resiliência das cadeias agroalimentares e à análise e desenho de melhores seguros agrícolas do ponto de vista do agricultor.

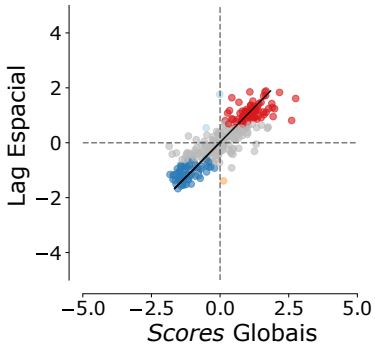
O clima desempenha um papel fundamental na produção de tomate. Isso leva a impactos

Figura 75 – Moran *Scartterplot* dos *Scores* Globais das soluções de referência Intermediária.



Fonte: Elaborada pelo autor.

Figura 76 – Moran *Scartterplot* dos *Scores* Globais das soluções de referência Superior.



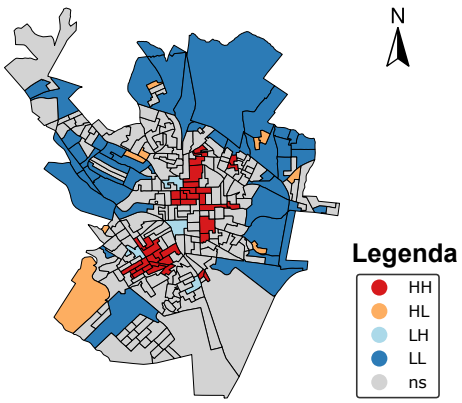
Fonte: Elaborada pelo autor.

diretos como as perdas anuais significativas relacionadas a perigos como granizo e geada e indiretos como o aumento da ocorrência de pragas. Adicionalmente, a produção de tomate apresenta processos de produção tradicionais e em estufa, e zonas de produção distribuídas por todo o país.

5.3.1 Área de estudo

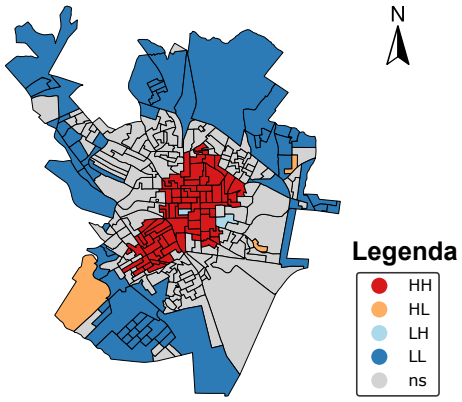
O presente estudo de caso envolveu quatro variáveis coletadas de 2006 a 2019 para todas as apólices de seguro agrícola para produção de tomate de vinte e sete cidades do sul do estado de São Paulo (Figura 82, escolhidas a partir de entrevistas realizadas com o grupo de especialistas considerando: (i) localização da fazenda (cidade); (ii) prêmio pago pelo agricultor pela apólice de seguro; (iii) valor total segurado; e (iv) produtividade estimada. Seguindo o trabalho de (SILVA *et al.*, 2021), todas as variáveis foram divididas pela área segurada para permitir uma melhor comparação entre diferentes processos produtivos e tamanhos de fazendas.

Figura 77 – LISA *cluster* dos *Scores* Globais das soluções de referência Inferior.



Fonte: Elaborada pelo autor.

Figura 78 – LISA *cluster* dos *Score* Globais das soluções de referência Intermediária.



Fonte: Elaborada pelo autor.

5.3.2 Método AHP

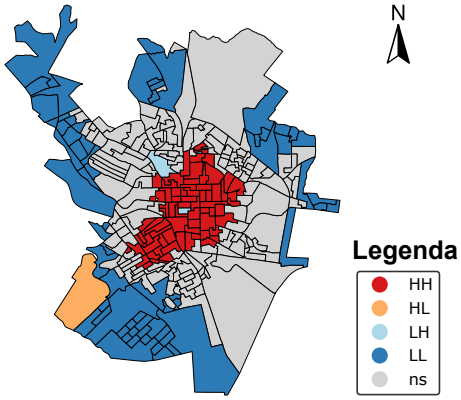
Uma vez definida as variáveis, camadas temáticas que serão utilizadas, foi utilizado o método AHP para definição de uma solução a partir da visão do grupo de especialista consultado, solicitando que fizessem a comparação pareada entre as variáveis a partir da escala fundamental de Saaty (1978). Após a coleta das respostas dadas, foi considerado como resultado a comparação que mais se repetia entre os especialistas e, a partir de então, foi utilizado a implementação do método AHP disponível no MultiMapas para a definição dos pesos.

A Tabela 14 apresenta a matriz de comparação pareada e os respectivos pesos de cada variável.

Uma vez com a matriz de comparação pareada das variáveis, foi calculada a CR para validar as comparações pareadas entre variáveis, sendo essa igual a 0.05 ($CR = 0.05$), mostrando que os valores de prioridade relativa são consistentes.

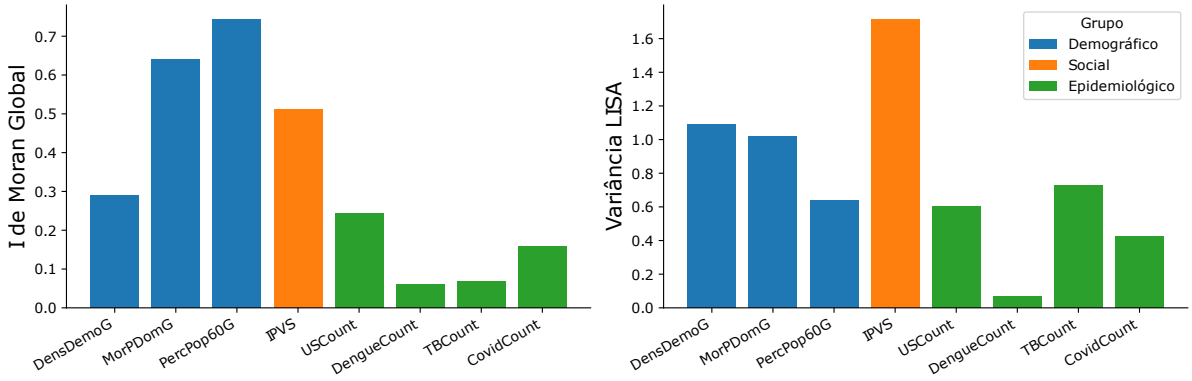
A partir desses resultados (Tabela 14), tem-se que a variável Produtividade Estimada

Figura 79 – LISA *cluster* dos *Scores* Globais das soluções de referência Superior.



Fonte: Elaborada pelo autor.

Figura 80 – Índice *I* de Moran Global e a variância do índice de Moran local, LISA, das camadas temáticas consideradas pelo GIS-moGA.



Fonte: Elaborada pelo autor.

(*PROD_EST*) foi a considerada mais importante entre as demais variáveis, seguida por Prêmio Líquido Área (*PRE_LIQ_AREA*), Limite Garantia Área (*LIM_GAR_AREA*) e Subvenção Federal Área (*SUB_FED_AREA*).

Com a definição do peso de cada variável é calculado um *score* global cruzando as diferentes camadas que representam cada variável ponderada pelos respectivos pesos que usam uma combinação linear ponderada (WLC), expressa pela Equação 5.4, em que a camada temática de adequação final é derivada multiplicando cada camada temática por seu peso relativo seguido pela soma dos resultados.

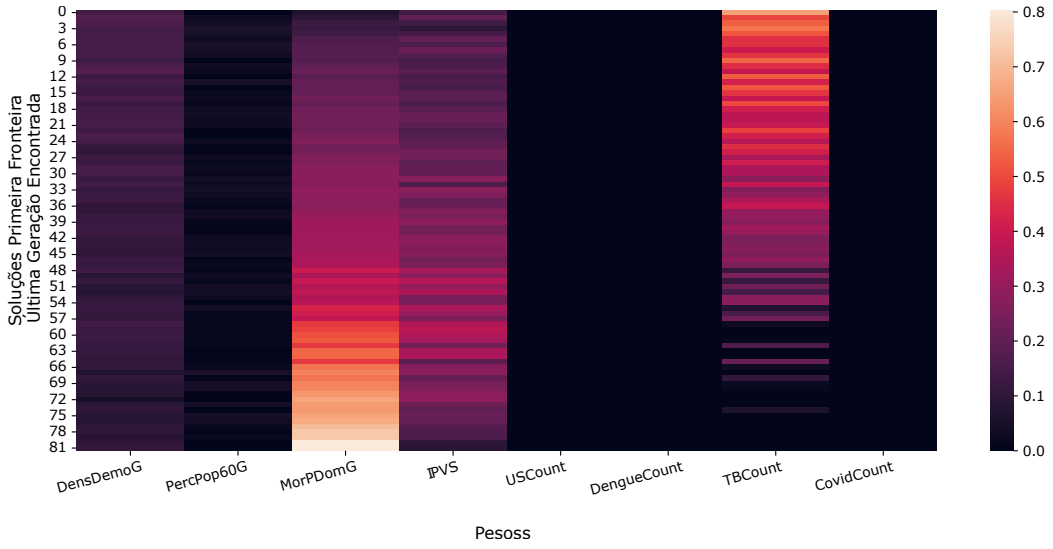
$$\mu_i = \begin{array}{rcll} 0.548 & \times & PROD\_EST_i & + \\ 0.283 & \times & PRE\_LIQ\_AREA_i & + \\ 0.117 & \times & LIM\_GAR\_AREA_i & + \\ 0.052 & \times & SUB\_FED\_AREA_i & \end{array} \tag{5.4}$$

Tabela 13 – Pesos dados pelos GIS-moGA nas soluções de referência.

Variável	Nome Variável	Pesos Solução de Referência Inferior	Pesos Solução de Referência Intermediária	Pesos Solução de Referência Superior
Densidade demográfica	*DensDemoG*	0,144	0,148	0,108
Média de moradores	MorPDomG	0,000	0,048	0,000
Percentual população acima de 60 anos	*PercPop*60*G*	0,066	0,137	0,804
IPVS	*IPVS*	0,142	0,147	0,088
Presença Unidade de Saúde	*HCCScore*	0,000	0,000	0,000
Casos de dengue	*DengueCount*	0,000	0,000	0,000
Caoso de tuberculose	*TbCount*	0,648	0,521	0,000
Caoso de COVID-19	*CovidCount*	0,000	0,000	0,000

Fonte: Dados da pesquisa.

Figura 81 – *Heatmap* dos pesos encontrado pelo GIS-moGA das variáveis na primeira fronteira de Pareto da última geração.

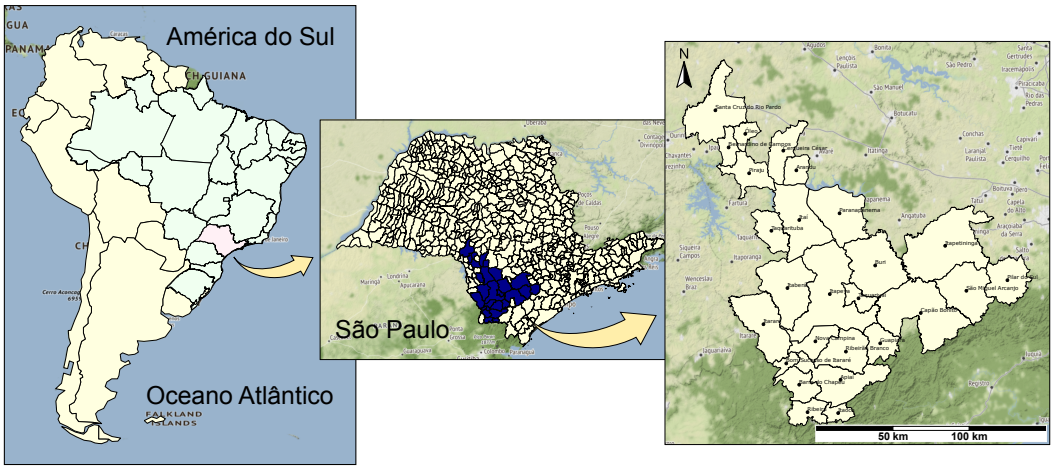


Fonte: Elaborada pelo autor.

A partir dos *scores* globais calculados, foi realizada a primeira análise para verificar suas estatísticas descritivas. Essas são mostradas na Tabela 15.

Os valores globais *scores* foram então divididos em cinco classes de acordo com o algoritmo *Natural Breaks*. Nesse caso, o método de classificação adotado tem a característica de agrupar valores semelhantes e maximizar as diferenças entre as classes com limites estabelecidos nos quais existem diferenças consideráveis entre os valores dos dados. Assim, esse método representa o escalonamento natural das séries de dados, agrupando-as de acordo com (MATSUMOTO; CATÃO; GUIMARÃES, 2017) similaridade. A Figura 83 apresenta a classificação dos setores

Figura 82 – Cidades da região de São Paulo escolhidas.



Fonte: Elaborada pelo autor.

Tabela 14 – Matriz de comparação pareada para as variáveis utilizadas no estudo de caso.

Variável	Nome Variável	Matriz Pareada	A	B	C	D	Pesos
ESTIMATED PRODUCTIVITY	*PROD_EST*	A	1.00	3.00	5.00	7.00	0.548
AREA PRIZE	*PRE_LIQ_AREA*	B	0.33	1.00	3.00	7.00	0.283
AREA WARRANTY LIMIT	*LIM_GAR_AREA*	C	0.20	0.33	1.00	3	0.117
FEDERAL SUBSIDITY AREA	*SUB_FED_AREA*	D	0.11	0.13	6	1.00	0.052

Fonte: Dados da pesquisa.

Tabela 15 – Estatísticas descritivas dos *scores* globais encontrado pelo método AHP.

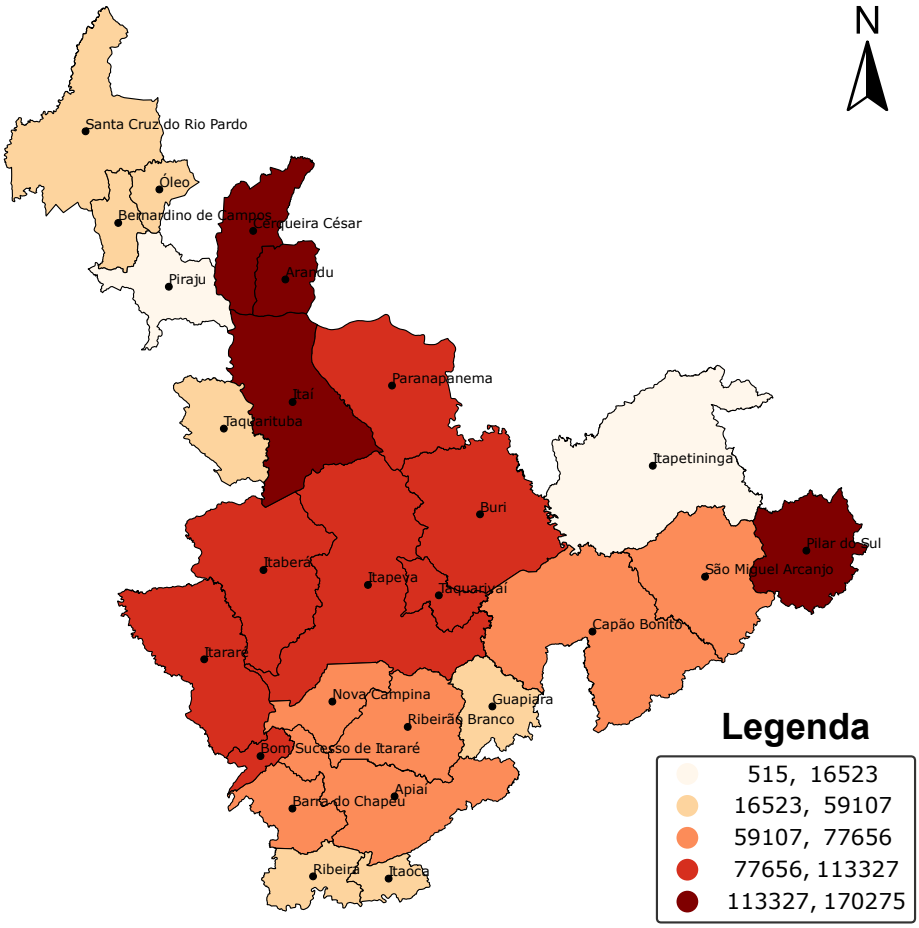
	Score Global
Média	3,07
Desvio Padrão	0,79
Mínimo	1,05
25%	2,56
50%	3,13
75%	3,62
Máximo	6,60

Fonte: Dados da pesquisa.

censitários em relação aos *escores* globais do modelo.

Para se certificar que a modelagem gerada pelo método AHP representa um fenômeno do ponto de vista espacial, e assim validarmos o mapeamento em questão, foi calculado o índice global *I* de Moran. Na Figura 84, pode-se verificar o valor de 0,23 do índice *I* de Moran global, o que há uma probabilidade inferior a 1% de que esse padrão de agrupamento possa

Figura 83 – Representação coroplética do *score* global gerado pelo método AHP das cidades da região de São Paulo escolhidas utilizando o algoritmo *Natural Breaks*.



Fonte: Elaborada pelo autor.

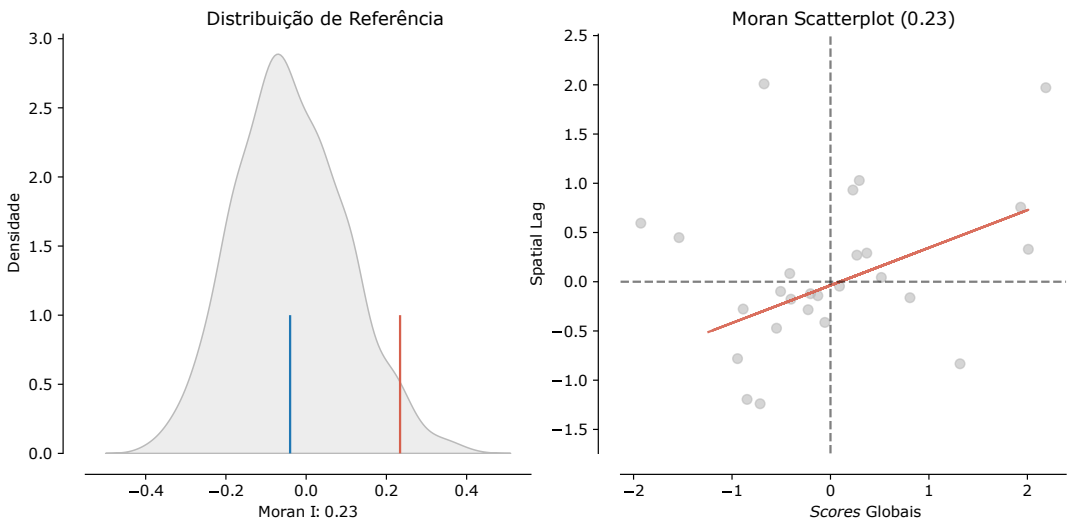
ser um resultado aleatório. Dessa forma, a distribuição espacial de valores altos e/ou baixos, no conjunto de dados, é mais espacialmente agrupada do que seria esperado se os processos espaciais subjacentes fossem aleatórios.

A partir do mapa temático modelados pelo método AHP (Figura 83), é possível observar que temos as cidades Cerqueirra César, Arandu, Itaí e Pilar do Sul são as cidades que concentram os maiores valores globais, indicando que essas são cidades-chave quanto aos valores prêmio, pagos pelo agricultor pela apólice de seguro e valor total segurado.

5.3.3 GIS-moGA

Apesar da modelagem do método AHP representar um fenômeno do ponto de vista espacial, foi também utilizado do GIS-moGA para gerar um *trade-off* de Pareto entre a maximização do índice *I* de Moran (dependência espacial) e a minimização do índice local de Moran LISA (heterogeneidade espacial) para os dados desse estudo de caso, considerando a definição de pesos

Figura 84 – Resultado do teste *I* de Moran dos *scores* globais gerados pelo método AHP.



Fonte: Elaborada pelo autor.

dada pela Equação 5.5.

$$\mu_i = \begin{array}{rcl} w_1 & \times & PROD\_EST \quad + \\ w_2 & \times & PRE\_LIQ\_AREA \quad + \\ w_3 & \times & LIM\_GAR\_AREA \quad + \\ w_4 & \times & SUB\_FED\_AREA \end{array} \tag{5.5}$$

A Figura 85 mostra os resultados da primeira fronteira encontrada pelo GIS-moGA na última geração do experimento, juntamente com as três soluções de referência e a solução gerada pelo método AHP (solução do especialista).

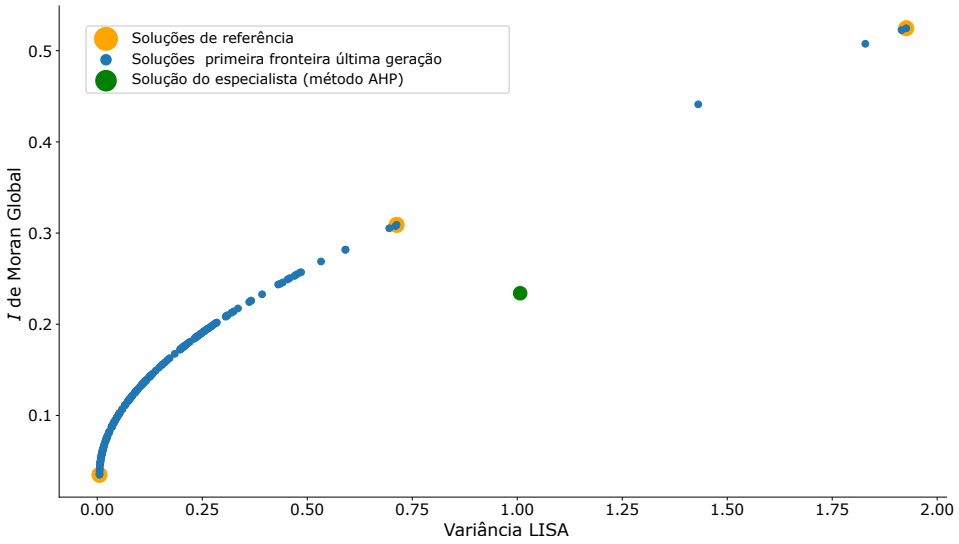
A partir das soluções de referência, foi realizada a determinação do algoritmo que melhor descreve número de classes e os limites de cada classe da distribuição do *score global* para construção da representação coroplética. Essa escolha foi realizada a partir do cálculo do ADCM do *score global*como pode ser observado nas Figuras 86, 87 e 88 de cada solução de referência. Vê-se, dessa forma, que ,o menor ADCM em cada solução de referência encontrado foi o algoritmo *Fisher Jenks*.

Uma vez determinado o algoritmo que melhor descreve o *score global* de cada solução de referência, pode-se analisar a distribuição desses valores. Na Tabela 16, tem-se as estatísticas descritivas dos *scores globais* de cada solução de referência.

As Figuras 89, 90 e 91 apresentam as representações coropléticas dos *scores* globais de cada solução de referência, segundo o algoritmo de Fisher Jenks.

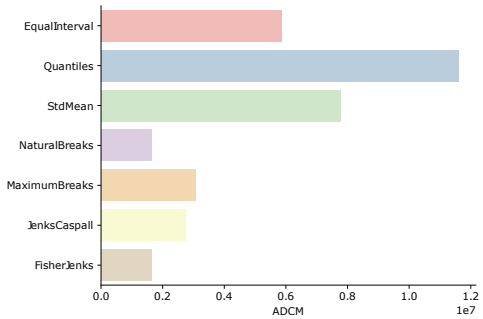
Nas Figuras 92, 93 e 94 temos o Moran *Scatterplot* que mostra o *lag* espacial do *score global* das soluções de referência. É possível observar que a solução de maximização do *I* de

Figura 85 – Soluções da primeira fronteira encontrada pelo GIS-moGA na última geração.



Fonte: Elaborada pelo autor.

Figura 86 – ADCM do *Score Global* das soluções de referência Inferior.



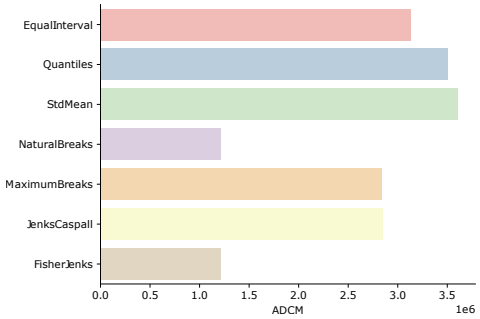
Fonte: Elaborada pelo autor.

Moran Global apresenta um número mais significativo de regiões com associações lineares espaciais alto-alto (HH) e baixo-baixo (LL), com diminuição desses tipos de associações para a solução intermediária e minimização da variância LISA.

As Figuras 95, 96 e 97 mostram os mapas com os *clusters* LISA das soluções de referência. É possível notar que na solução de referência inferior, que minimiza o LIZA, apresenta apenas dois *clusters* um HH entre a cidade Cerqueira César e outro LL entre as cidades Cruz do Rio Pardo, Óleo e Bernardino de Campos, já as soluções de referência intermediária e superior apresentaram o mesmo *clusters*, sendo que o *cluster* LL é o mesmo da solução de referência inferior, já o *cluster* HH, além da cidade Cerqueira César que apareceu na solução de referência inferior, também apareceram as cidades Arandu e Itaí.

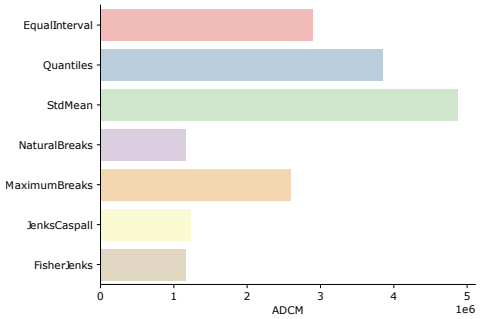
Para verificar-se qual variável contribuiu mais para a dependência e heterogeneidade espacial, primeiramente analisaram-se os valores do *I* de Moran Global e a variância do índice

Figura 87 – ADCM do *Score Global* das soluções de referência Intermediária.



Fonte: Elaborada pelo autor.

Figura 88 – ADCM do *Score Global* das soluções de referência Superior.



Fonte: Elaborada pelo autor.

Tabela 16 – Estatísticas Descritivas das Soluções de Referência.

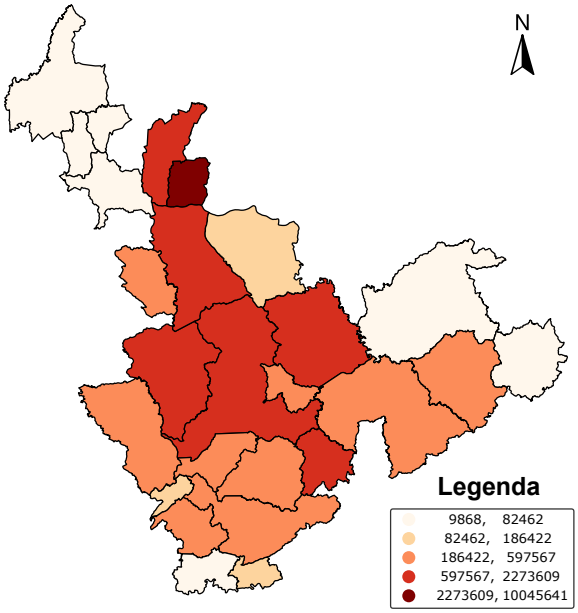
	Score Global Solução de Referência Inferior	**Score Global Solução de Referência Intermediária**	**Score Global Solução de Referência Superior**
Média	806783,92	437058,08	504692,65
Desvio Padrão	1945451,29	630585,53	682648,16
Mínimo	9867,70	3363,67	3866,70
25%	94121,33	145339,33	51728,38
50%	338204,28	248323,99	271473,00
75%	664692,49	458325,33	522587,04
Máximo	10045641,22	3052465,35	2413232,10

Fonte: Dados da pesquisa.

de Moran local, LISA, de cada uma das variáveis. A Figura 98 apresenta o valor do índice *I* de Moran Global e a variância do índice de Moran local, LISA, das variáveis que foram consideradas pelo GIS-moGA no processo de otimização.

Na Tabela 17, temos os pesos, o grau de importância, dados pelo GIS-moGA para cada variáveis de referência.

Figura 89 – Representação coroplético do *Score Global* segundo o algoritmo Fisher Jenks das soluções de referência Inferior.



Fonte: Elaborada pelo autor.

Tabela 17 – Pesos dados pelos GIS-moGA nas soluções de referência.

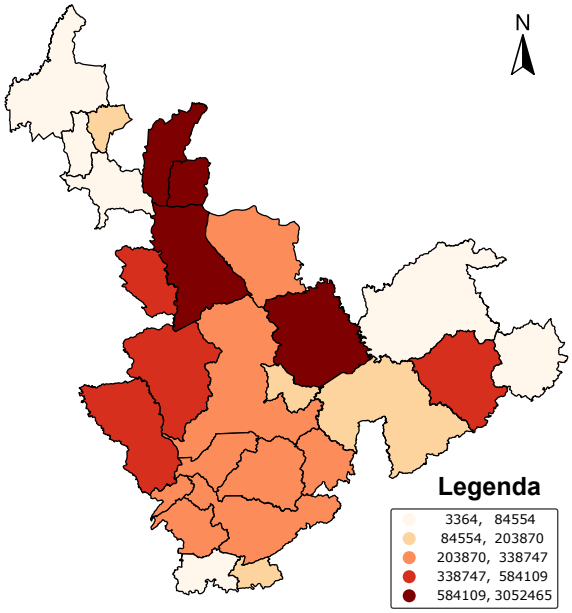
Variável	**Nome Variável**	**Pesos Solução de Referência Inferior**	**Pesos Solução de Referência Intermediária**	**Pesos Solução de Referência Superior**
Produtividade Estimada	*PROD_EST*	0.077	0.023	0.112
Prêmio Líquido	*PRE_LIQ_AREA*	0.321	0.464	0.788
Limite Garantia	*LIM_GAR_AREA*	0.434	0.119	0.088
Subvenção Federal	*SUB_FED_AREA*	0.168	0.393	0.012

Fonte: Dados da pesquisa.

Analisando os dados da Tabela 17, é possível notar que a variável *PRE_LIQ_AREA* foi a que mais contribuiu para a dependência e heterogeneidade espacial global nas soluções de referência intermediária e superior de acordo com o GIS-moGA

Na Figura 99, temos um *heatmap* dos pesos encontrados pelo GIS-moGA das variáveis na primeira fronteira de Pareto da última geração. É possível observar que a variável *PRE_LIQ_AREA* é a que mais contribui, justamente pelo fato dela ser a que possui o maior índice *I* de Moran global, apesar de possuir uma variância elevada em comparação com as demais. Também pode-se destacar a variável *LIM_GAR_AREA* contribuiu, uma vez que é a que possui a menor variância entre as demais.

Figura 90 – Representação coroplético do *Score Global* segundo o algoritmo Fisher Jenks das soluções de referência Intermediária.



Fonte: Elaborada pelo autor.

5.4 Estudo de caso 3: estatística de varredura espacial

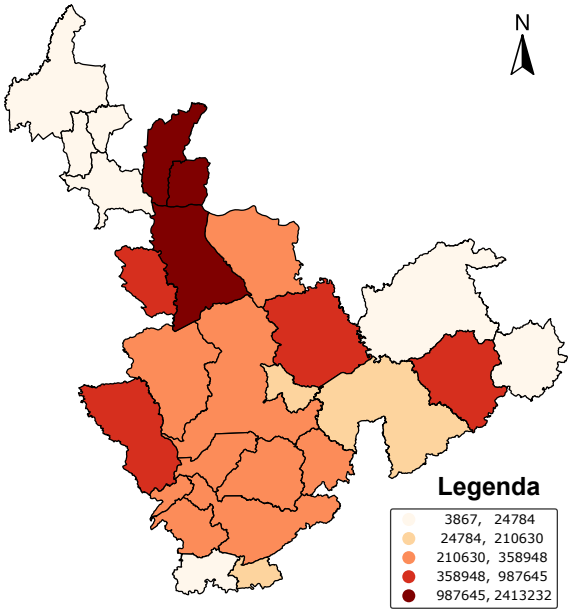
A técnica estatística de varredura espacial foi desenvolvida por (KULLDORFF; NAGARWALLA, 1995), conforme já mencionado, visando identificar e localizar *clusters* de risco presentes em uma determinada região de estudo. Para essa análise, foram utilizados todos os casos notificados e confirmados de dengue do Sistema de Informação de Agravos de Notificação-Sinan Dengue/Chikungunya de moradores da zona urbana do município de São Carlos-SP, no período de 1º de janeiro a 31 de dezembro dos anos de 2018, 2019 e 2020, além do IPVS, calculado para cada setor censitário (<https://ipvs.seade.gov.br/view/index.php>)

Os *clusters* de risco são identificados graficamente por janelas circulares de raio variável ao redor dos centroides de cada setor censitário, para as quais é calculado o número esperado de ocorrências dentro do círculo. A região delimitada pela janela de análise, denominada região z, pode constituir um cluster caso o valor encontrado seja maior ou menor do que o esperado. Esse procedimento é realizado em todos os centroides em análise (LUCENA; MORAES, 2012).

Assim, a hipótese nula (H_0) contra a hipótese alternativa (H_1) foi testada de forma diferente entre as doenças, destacando que H_0 assume que não há aglomerados de casos de dengue, ou seja, a população tem a mesma probabilidade de contrair caso de dengue e H_1 assume que uma ou mais regiões z são áreas em que haveria maior ou menor probabilidade de contrair as doenças, em comparação com a que estão fora dessa área.

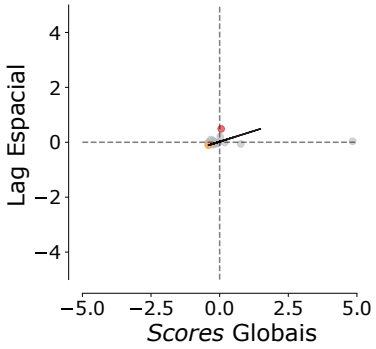
A fim de identificar *clusters* puramente espaciais em que a distribuição é heterogênea. Os eventos são raros em relação à população. Foi utilizado o modelo discreto de Poisson com

Figura 91 – Representação coroplético do *Score Global* segundo o algoritmo Fisher Jenks das soluções de referência Superior.



Fonte: Elaborada pelo autor.

Figura 92 – Moran *Scartterplot* do *Score Global* das soluções de referência Inferior.

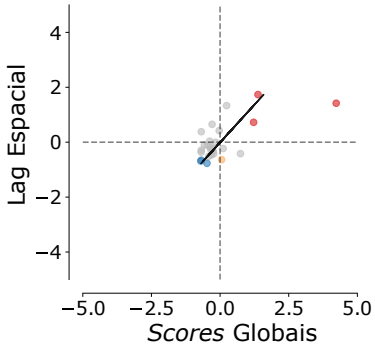


Fonte: Elaborada pelo autor.

requisitos de *clusters* geográficos não sobrepostos, *clusters* com formato circular, 999 replicações e tamanho da população exposta estipulado pelo coeficiente de Gini divulgado pelo próprio software. Nesse modelo, o número de casos foi comparado aos dados populacionais da linha de base e o número esperado de casos, de cada unidade, foi proporcional à população em risco (ALVES *et al.*, 2019).

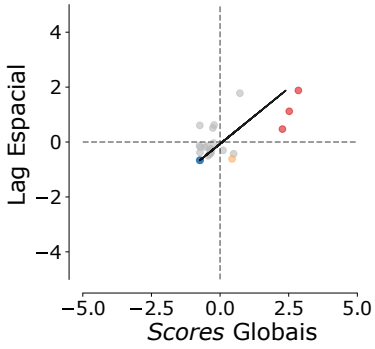
O risco relativo (RR) de cada *cluster* foi calculado, permitindo a comparação de informações em diferentes áreas, indicando a intensidade de ocorrência de casos de dengue no município de São Carlos. Vale ressaltar que o *RR* será definido como o risco de ter dengue em uma área de risco do município em relação ao risco de ter dengue fora dessa área.

Figura 93 – Moran *Scartterplot* do *Score Global* das soluções de referência Intermediária.



Fonte: Elaborada pelo autor.

Figura 94 – Moran *Scartterplot* do *Score Global* das soluções de referência Superior.



Fonte: Elaborada pelo autor.

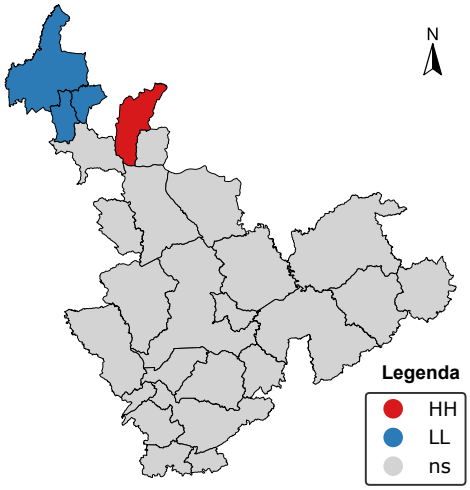
Áreas com $p\text{-}valor < 0,05$ foram consideradas estatisticamente significativas. O intervalo de confiança foi calculado e estimado em 95% (ARROYO *et al.*, 2017). A identificação dos *RR* dos *clusters* permite a comparação de informações em áreas distintas, pois desconsideram-se os efeitos de diferentes populações, resultando na intensidade de ocorrência do fenômeno sob análise ao longo da área de estudo. Os valores resultantes desse cálculo serão denominados de alto risco quando o *cluster RR* for maior que um ($RR > 1$), e baixo risco quando menor que um ($RR < 1$) (ARROYO *et al.*, 2017).

As análises de detecção de agrupamento foram realizadas usando o *software* SaTScan, um *software* gratuito amplamente utilizado por importantes Centros de Estudos em Saúde (ALVES *et al.*, 2019), na versão 10.1.

O Sinan Dengue/Chikungunya registrou 762, 10.451 e 1.650 casos de dengue em 2018, 2019 e 2020, respectivamente, no município de São Carlos.

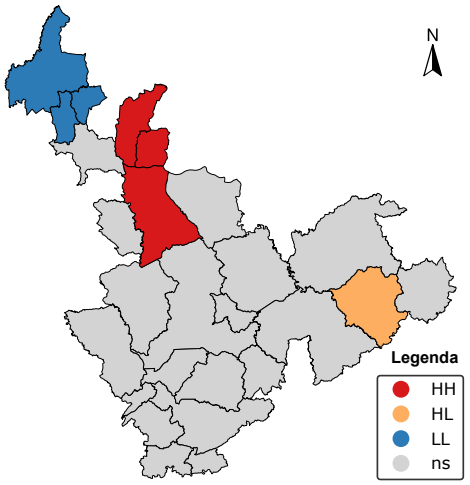
Aplicando as estatísticas de varredura espacial para casos de dengue, três *clusters* estatisticamente significativos foram detectados em 2018, vinte *clusters* estatisticamente significativos em 2019 e cinco *clusters* estatisticamente significativos em 2020, como pode ser visto na Fi-

Figura 95 – LISA *cluster* do *Score Global* das soluções de referência Inferior.



Fonte: Elaborada pelo autor.

Figura 96 – LISA *cluster* do *Score Global* das soluções de referência Intermediária.



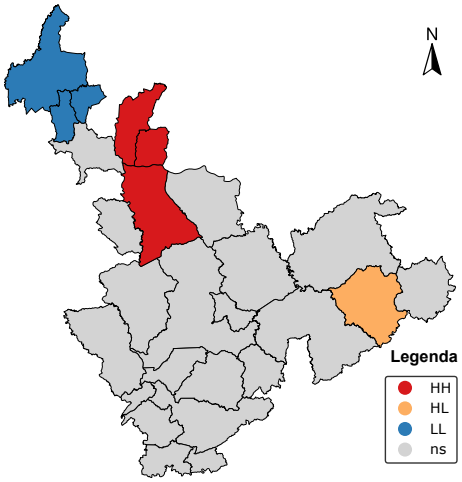
Fonte: Elaborada pelo autor.

gura 100, Figura 101 e Figura 102, respectivamente. Tabela 18, Tabela 19 e Tabela 20 mostram as características dos três clusters estatisticamente significativos com maior risco para dengue, segundo varredura temporal, no município de São Carlos-SP nos anos de 2018, 2019 e 2020, respectivamente.

O município de São Carlos, assim como outros polos industriais do Brasil, sofreu significativa urbanização em suas regiões periféricas (LIMA, 2007). Parte significativa desse crescimento ocorreu na forma de loteamentos de padrão precário e concentrados na população de baixa renda, nos setores sudeste e sul da cidade (SCHENK; FANTIN; PERES, 2015).

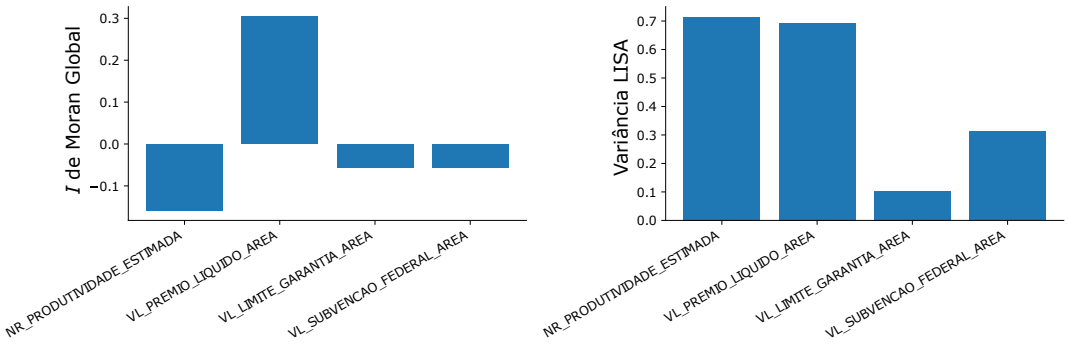
Ao observar a Figura 100, Figura 101 e Figura 102, é possível detectar que os setores censitários que compuseram os clusters de risco puramente espacial são oriundos do bairro

Figura 97 – LISA *cluster* do *Score Global* das soluções de referência Superior.



Fonte: Elaborada pelo autor.

Figura 98 – Índice *I* de Moran Global e a variância do índice de Moran local, LISA, das camadas temáticas consideradas pelo GIS-moGA.



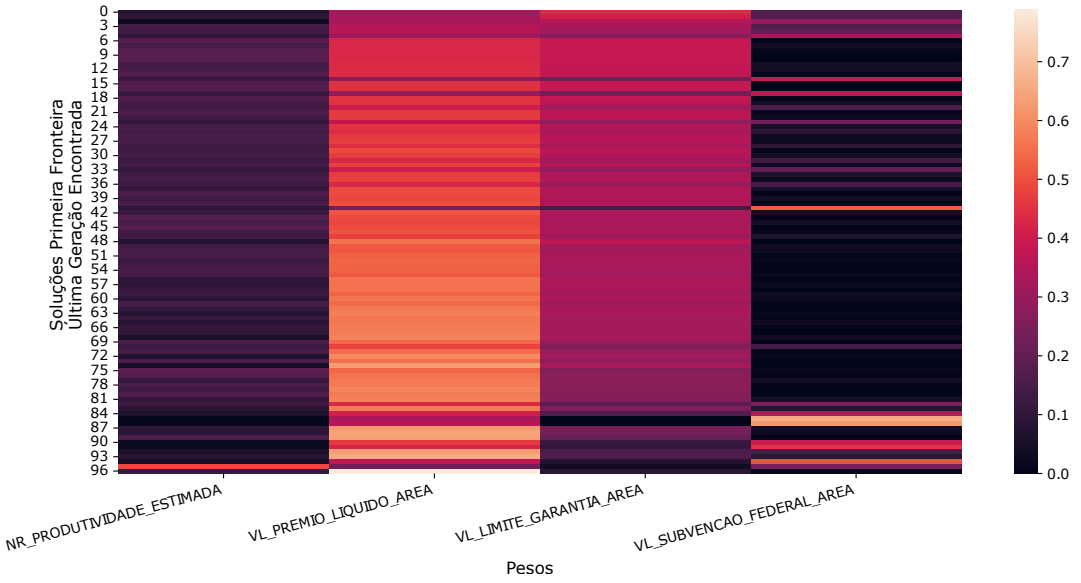
Fonte: Elaborada pelo autor.

Cidade Aracy, também localizado na zona sul da cidade, e do bairro Jardim Tangará, esse na região nordeste de São Carlos, região com alto índice social de vulnerabilidade, de acordo com o IPVS.

Por outro lado, a presente análise resultou em um *cluster* de baixo risco relativo para casos de dengue, sendo esses setores censitários localizados nos bairros Vila Prado, Vila Boa Vista, Vila Carmem e Jardim Beatriz, localizados na região sudoeste do município. Essas localidades apresentaram taxa de incidência abaixo da média, ou seja, o número de casos espaciais foi menor do que em qualquer outra região do município, constituindo-se em áreas de proteção para infecção por casos de dengue.

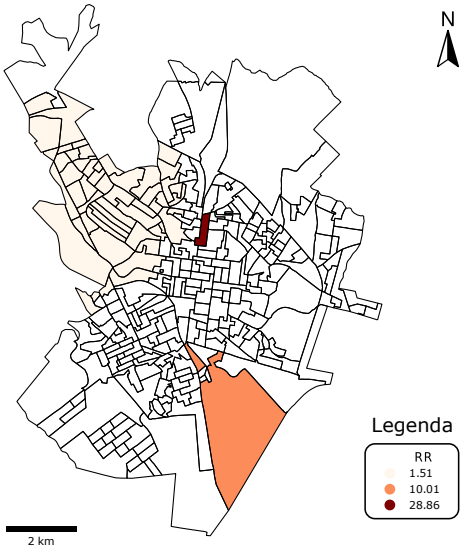
Vale ressaltar que o município de São Carlos é considerado polo de alta tecnologia, referência para o estado de São Paulo, com IDH, índice de Gini e incidência de pobreza melhores que a média estadual, o que pode explicar a ausência de setores censitários como vulnerabilidade

Figura 99 – *Heatmap* dos pesos encontrado pelo GIS-moGA das variáveis na primeira fronteira de Pareto da última geração.



Fonte: Elaborada pelo autor.

Figura 100 – Aglomerados espaciais da análise da estatística de varredura espacial dos casos de dengue em 2018.

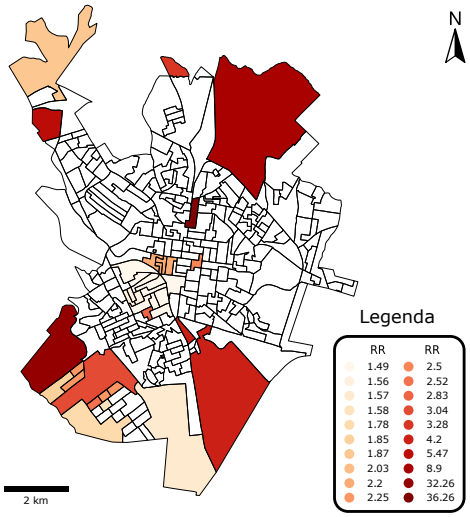


Fonte: Elaborada pelo autor.

muito alta.

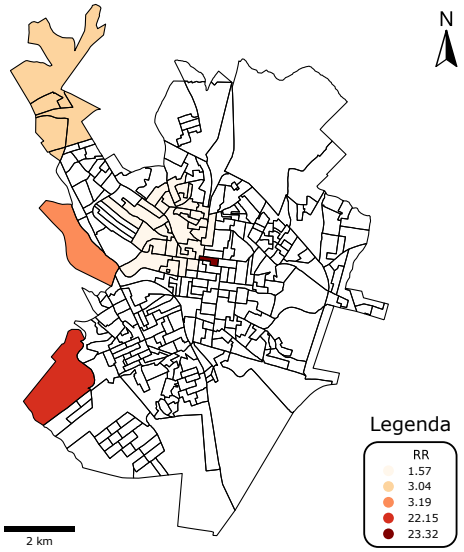
Como limitação desse estudo de caso, destaca-se, que em estudos ecológicos, os resultados identificados não podem ser interpretados ao nível individual. Sem dúvida, as estatísticas de varredura espacial contribuíram para expor o cenário da dengue em São Carlos e a presença de áreas geográficas, do município, mais susceptíveis ao adoecimento e que necessitam de ações

Figura 101 – Aglomerados espaciais da análise da estatística de varredura espacial dos casos de dengue em 2019.



Fonte: Elaborada pelo autor.

Figura 102 – Aglomerados espaciais da análise da estatística de varredura espacial dos casos de dengue em 2020.



Fonte: Elaborada pelo autor.

específicas para o controle da doença.

5.5 Considerações Finais

Neste capítulo, foi descrito três estudos de casos de aplicação do MultiMapas, que é a principal contribuição desta tese. Percebeu-se que o MultiMapas pode ser utilizado em diferentes contextos, áreas e fontes de dados com diferentes aspectos espaciais.

Tabela 18 – Características dos aglomerados estatisticamente significativos no município de São Carlos em 2018.

	Cluster 1	Cluster 2	Cluster 3
Número de setores censitários	1	1	52
População	553	527	42612
Número de casos	45	17	194
Número de casos esperados	1.66	1.74	142.01
RR	28.86	10.01	1.51

Fonte: Dados da pesquisa.

Tabela 19 – Características dos aglomerados estatisticamente significativos no município de São Carlos em 2019.

	Cluster 1	Cluster 2	Cluster 3
Número de setores censitários	1	1	1
População	553	138	80
Número de casos	831	192	33
Número de casos esperados	24.97	6.07	3.72
RR	36.26	32.26	8.90

Fonte: Dados da pesquisa.

Tabela 20 – Características dos aglomerados estatisticamente significativos no município de São Carlos em 2020.

	Cluster 1	Cluster 2	Cluster 3
Número de setores censitários	1	1	1
População	597	138	911
Número de casos	84	20	19
Números casos esperados	3.82	0.92	6.01
RR	23.32	22.15	3.19

Fonte: Dados da pesquisa.

CAPÍTULO

6

CONCLUSÕES

A tomada de decisão para problemas complexos com base em fontes de dados heterogêneas e múltiplas requer a estruturação de informações com representação adequada ao fenômeno em análise. Dimensões como a espacial adicionam complexidade ao processamento de dados, extração de informações e interpretação de resultado. Esta tese apresentou o MultiMapas, uma metodologia que coleta, analisa e disponibiliza dados com características espaciais de maneira eficiente e eficaz.

O MultiMapas foi avaliado a partir de três estudos de casos distintos, demonstrando a sua eficácia.

6.1 Contribuições

As principais contribuições desta tese são:

- Uma metodologia para análise de dados com características espaciais que: (1) Realiza a carga de dados de múltiplas e heterogêneas fontes, anonimamente; (2) Geolocalização e Agregação de dados, considerando a granularidade espacial da área de estudo; (3) Definição de camadas temáticas com propriedades de dados e representação espacial de interesse relacionados ao fenômeno;

- Implementação de uma técnica de análise multicritério para definir a influência de cada camada temática para criar um mapa coroplético temático com o padrão espacial do fenômeno através da álgebra de mapas;

- Implementação de um algoritmo genético multiobjetivo que otimiza a dependência e heterogeneidade espacial;

- Uma ferramenta de código aberto que integra consistentemente dados de diferentes fontes e contextos, otimizando e facilitando a análise, gestão ou representação do espaço e dos fenômenos que nele ocorrem por profissionais de diferentes áreas.

6.2 Publicações

Durante o período do doutorado foram publicados os seguintes artigos em conferências, periódicos e capítulos de livros:

- LOPES, G. R., DELBEM, A. C. B. (2020). **MultiMaps geo-referenced spatiotemporal analyzes of multiple epidemics**. IV Encontro Paulista de Pós-Graduandos em Computação.

- LOPES, G. R., DELBEM, A. C. B., SOUSA, J. B. (2021). **Introdução à análise de dados geoespaciais com python**. Minicursos do XIV Encontro Unificado de Computação do Piauí (ENUCOMPI) e XI Simpósio de Sistemas de Informação (SINFO). Sociedade Brasileira de Computação. Sociedade Brasileira de Computação. DOI: <https://doi.org/10.5753/sbc.7669.6.5>

- LOPES, G. R., PELARIGO, K. J., DELBEM, A. C. B., SOUSA, J. B. (2022). **Análise Exploratória de Dados Espaciais com Python**. Minicursos da X Escola Regional de Computação do Ceará, Maranhão e Piauí Sociedade Brasileira de Computação. DOI:<https://doi.org/10.5753/sbc.11014.1.1>

- LOPES, G., PELARIGO, K., DELBEM, A. C. B. (2022). **Identification of risk areas as a method of surveillance of dengue cases**. Anais da X Escola Regional de Computação do Ceará, Maranhão e Piauí, (pp. 51-60). Porto Alegre: SBC. DOI: <https://doi.org/10.5753/ercemapi.2022.225892>

- LOPES, G. R., DELBEM, A. C. B., SILVA, R. F., JÚNIOR, C. B., DE MATTOS, S. H. V. L., SCATOLINI, D., GHIGLIENO, F., SARAIVA, A. M. (2022). **MultiMaps: a tool for decision-making support in the analyzes of multiple epidemics**. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Spatial Computing for Epidemiology (pp. 22-25). DOI: <https://doi.org/10.1145/3557995.3566119>

- LOPES, G. R., SILVA, R. F., PELARIGO, K. J., YAMAMURA, M., DELBEM, A. C. B.; SARAIVA, A. M. (2023). **Identification of risk areas using spatial clustering to improve dengue monitoring in urban environments**. Revista de Sistemas e Computação-RSC, 12(3).DOI: <https://doi.org/10.36558/rsc.v12i3.7921>

- LOPES, G. R., SILVA, R. F., PELARIGO, K. J., YAMAMURA, M., DELBEM, A. C. B., SCATOLINI, D., GHIGLIENO, F., Saraiva, A. M. (2023). **Proposal of a framework for**

improving multi-criteria decision-making related to epidemics using heterogeneous spatial data and evolutionary algorithms. Research, Society and Development, 12(2), e0212239844-e0212239844. DOI: <https://doi.org/10.33448/rsd-v12i2.39844>

- LOPES, G. R., DELBEM, A. C. B., SILVA, R. F., JÚNIOR, C. B., DE MATTOS, S. H. V. L., SCATOLINI, D., SARAIVA, A. M. (2023). **Use of multicriteria analysis and map algebra to identify risk areas for multiple health aggravations.**. LV Simpósio Brasileiro de Pesquisa Operacional.

- YAMAMURA, M., ABRANHÃO, R. M. C. de M., PALASIO, R. G. S., SILVA, B. P. M. da S., LOPES, G. R., NETO, F. C. (2023). **Tuberculose e COVID-19: distribuição espacial e áreas de coocorrência em Araraquara/SP (2020-2022)**. 58º Congresso da Sociedade Brasileira de Medicina Tropical (MEDTROP 2023).

6.3 Dificuldades

A tomada de decisão para problemas complexos a partir de dados com características espaciais é extremamente desafiadora e envolve esforços de diversas áreas.

Neste contexto, uma das maiores dificuldades encontradas para o desenvolvimento do trabalho apresentado foi o acesso a bases de dados com características espaciais. Em muitos casos, os dados apresentam restrições devido a questões de privacidade e ao ter acesso, muitas das vezes, esses dados apresentam inconsistências, erros e ruído, dificultando determinar a qualidade e corretude dos deles. Outro aspecto relacionado aos dados consiste na defasagem temporal dos dados, ou seja, refere-se ao atraso entre o momento em que os dados são gerados e o momento em que estão disponíveis para uso e processamento.

Outro aspecto fundamental foi de compreender os conceitos, as ferramentas, as metodologias de avaliação e o levantamento do estado da arte na integração de técnicas MCDM e GIS. O mapeamento do estado da arte demandou um longo período devido ao amplo volume de artigos publicados e a falta de conhecimento prévio, demandou esforço e tempo para que pudesse ser implementado corretamente.

A otimização da dependência e heterogeneidade espacial foi realizada através de algoritmos genéticos. O desenvolvimento de operadores genéticos, bem como, a calibração do número de indivíduos, número e populações, taxa de mutação e cruzamento também demandaram tempo e experimentos extensivos. As inúmeras repetições de experimentos em função do comportamento estocástico das meta-heurísticas produziram uma quantidade de dados razoavelmente significativas para tratar. *Scripts* foram desenvolvidos na tentativa de auxiliar no processamento das inúmeras avaliações.

6.4 Trabalhos Futuros

Esta tese não encerra as possibilidades de estudos relacionados ao desenvolvimento de métodos para integrar MCDM e GIS. Outras pesquisas devem ser conduzidas a partir dos resultados descritos neste trabalho. Dentre os próximos trabalhos, os seguintes tópicos merecem destaquem:

- **Aplicação dos métodos em outros estudos de caso:** Nese trabalho foi utilizado em apenas três estudos de casos. Os métodos propostos nesta tese devem ser adaptados, aplicados e avaliados em outros cenários;
- **Considerar aspectos temporais:** nos estudos de casos realizados nesta tese, foram considerados apenas o aspecto espacial. No entanto, o aspecto temporal é uma dimensão crucial para entender padrões e tendências em conjuntos de dados espaciais e uma investigação considerando o aspecto temporal pode ser realizada;
- **Operadores genéticos:** um dos alicerces da computação evolutiva é o conjunto de operadores genéticos (mutação e cruzamento). A pesquisa de operadores mais eficientes pode reduzir o tempo de convergência da otimização, encontrando soluções tão boas quanto às encontradas nesta tese;
- **Utilizar outros indicadores de dependência e heterogeneidade espacial:** sugerem-se utilizar outros indicadores de dependência e heterogeneidade espacial como as estatísticas de Geary e Getis-Ord.
- **Abordagem multi-objetivo:** o GIS-moGA apresenta uma implementação baseada no NSGA-II, sugere-se a comparação de outros métodos multiobjetivos como MOEAD-D, NSGA-III, PESA-II e VEGA.

REFERÊNCIAS

AJRINA, A. S.; SARNO, R.; GINARDI, H.; FAJAR, A. Mining zone determination of natural sandy gravel using fuzzy ahp and saw, moora and copras methods. **International Journal of Intelligent Engineering & Systems**, v. 13, n. 5, 2020. Citado nas páginas 80 e 85.

AKGÜN, İ.; ERDAL, H. Solving an ammunition distribution network design problem using multi-objective mathematical modeling, combined ahp-topsis, and gis. **Computers & Industrial Engineering**, Elsevier, v. 129, p. 512–528, 2019. Citado nas páginas 79 e 82.

ALMEIDA, E. Econometria espacial. **Campinas–SP. Alínea**, v. 31, 2012. Citado nas páginas 33, 34, 35, 36, 44, 45, 46, 47, 48, 50, 51, 52, 91 e 122.

ALVES, L. S.; SANTOS, D. T. D.; ARCOVERDE, M. A. M.; BERRA, T. Z.; ARROYO, L. H.; RAMOS, A. C. V.; ASSIS, I. S. D.; QUEIROZ, A. A. R. D.; ALONSO, J. B.; ALVES, J. D. *et al.* Detection of risk clusters for deaths due to tuberculosis specifically in areas of southern brazil where the disease was supposedly a non-problem. **BMC infectious diseases**, BioMed Central, v. 19, n. 1, p. 1–13, 2019. Citado nas páginas 138 e 139.

ANDRADE, A. L.; MONTEIRO, A. M. V.; BARCELLOS, C.; LISBOA, E.; ACOSTA, L. M. W.; ALMEIDA, M. C. d. M.; BRITO, M. R. V.; CARVALHO, M. S.; SANTOS, M. A. d.; CRUZ, O. *et al.* Introdução à estatística espacial para a saúde pública. 2007. Citado nas páginas 33, 34, 44, 45, 46 e 101.

ANSELIN, L. **Spatial econometrics: methods and models**. [S.l.]: Springer Science & Business Media, 1988. v. 4. Citado na página 44.

____. Local indicators of spatial association—lisa. **Geographical analysis**, Wiley Online Library, v. 27, n. 2, p. 93–115, 1995. Citado nas páginas 29 e 51.

____. Exploring spatial data with geodatm: a workbook. **Center for spatially integrated social science**, p. 165–223, 2005. Citado nas páginas 44 e 45.

ANSELIN, L.; FLORAX, R.; REY, S. J. **Advances in spatial econometrics: methodology, tools and applications**. [S.l.]: Springer Science & Business Media, 2013. Citado nas páginas 44 e 45.

ARROYO, L. H.; YAMAMURA, M.; PROTTI-ZANATTA, S. T.; FUSCO, A. P. B.; PALHA, P. F.; RAMOS, A. C. V.; UCHOA, S. A.; ARCÊNCIO, R. A. Identificação de áreas de risco para a transmissão da tuberculose no município de são carlos, são paulo, 2008 a 2013. **Epidemiologia e Serviços de Saúde**, SciELO Brasil, v. 26, p. 525–534, 2017. Citado na página 139.

BÄCK, T.; FOGEL, D. B.; MICHALEWICZ, Z. **Evolutionary computation 1: Basic algorithms and operators**. [S.l.]: CRC press, 2018. Citado na página 66.

BEIRIGO, B. A.; SANTOS, A. G. dos. Application of nsga-ii framework to the travel planning problem using real-world travel data. In: IEEE. **2016 IEEE Congress on Evolutionary Computation (CEC)**. [S.l.], 2016. p. 746–753. Citado na página 91.

BEYER, J.; HEESCHE, K.; HAUPTMANN, W.; OTTE, C.; KRUSE, R. Ensemble learning for multi-source information fusion. **Intelligent Autonomous Systems: Foundations and Applications**, Springer, p. 123–141, 2010. Citado na página 31.

BOSELA, R.; EISSA, M.; SHOUAKAR-STASH, O.; ALI, M. E.; SHAWKY, H. A.; SOLIMAN, E. A. Potential aquifer mapping for cost-effective groundwater reverse osmosis desalination in arid regions using integration of hydrochemistry, environmental isotopes and gis techniques. **Groundwater for Sustainable Development**, Elsevier, v. 19, p. 100853, 2022. Citado nas páginas 78 e 81.

BRAUERS, W. K. M.; ZAVADSKAS, E. K. Project management by multimoora as an instrument for transition economies. **Technological and economic development of economy**, Taylor & Francis, v. 16, n. 1, p. 5–24, 2010. Citado na página 84.

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. Citado nas páginas 33, 36, 37 e 117.

CAO, L.; MIAO, F.; YANG, W. Application of hgml-based fire data management in fire command system. **Journal of Theoretical & Applied Information Technology**, v. 47, n. 1, 2013. Citado na página 28.

CENSO, I. Disponível em:< http://www.censo2010.ibge.gov.br/>. **Acesso em**, v. 23, 2010. Citado nas páginas 47 e 102.

CHABUK, A.; AL-ANSARI, N.; HUSSAIN, H. M.; KNUTSSON, S.; PUSCH, R.; LAUE, J. Combining gis applications and method of multi-criteria decision-making (ahp) for landfill siting in al-hashimiyah qadhaa, babylon, iraq. **Sustainability**, MDPI, v. 9, n. 11, p. 1932, 2017. Citado na página 72.

CHAN, A. H.; KWOK, W.; DUFFY, V. G. Using ahp for determining priority in a safety management system. **Industrial Management & Data Systems**, Emerald Group Publishing Limited, v. 104, n. 5, p. 430–445, 2004. Citado na página 56.

CLEMEN, R. T.; REILLY, T. **Making hard decisions with DecisionTools**. [S.l.]: Cengage Learning, 2013. Citado na página 110.

COELLO, C. A. C. **Evolutionary algorithms for solving multi-objective problems**. [S.l.]: Springer, 2007. Citado na página 69.

CRESSIE, N. **Statistics for spatial data**. [S.l.]: John Wiley & Sons, 2015. Citado na página 33.

DANESH, G.; MONAVARI, S. M.; OMRANI, G. A.; KARBASI, A.; FARSAD, F. Compilation of a model for hazardous waste disposal site selection using gis-based multi-purpose decision-making models. **Environmental monitoring and assessment**, Springer, v. 191, p. 1–14, 2019. Citado nas páginas 78 e 81.

DARWIN, C. **On the origin of species, 1859**. [S.l.]: Routledge, 2004. Citado na página 61.

DAWSON, T.; SANDOVAL, J. O.; SAGAN, V.; CRAWFORD, T. A spatial analysis of the relationship between vegetation and poverty. **ISPRS International Journal of Geo-Information**, MDPI, v. 7, n. 3, p. 83, 2018. Citado na página 52.

DEB, K. **Multi-objective optimisation using evolutionary algorithms: an introduction**. [S.l.]: Springer, 2011. Citado nas páginas 65, 66, 69 e 120.

DEB, K.; AGRAWAL, S.; PRATAP, A.; MEYARIVAN, T. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In: SPRINGER. **Parallel Problem Solving from Nature PPSN VI: 6th International Conference Paris, France, September 18–20, 2000 Proceedings 6**. [S.l.], 2000. p. 849–858. Citado na página 67.

DEB, K.; DEB, K. Multi-objective optimization. In: **Search methodologies: Introductory tutorials in optimization and decision support techniques**. [S.l.]: Springer, 2013. p. 403–449. Citado na página 69.

DEB, K.; PRATAP, A.; AGARWAL, S.; MEYARIVAN, T. A fast and elitist multiobjective genetic algorithm: Nsga-ii. **IEEE transactions on evolutionary computation**, IEEE, v. 6, n. 2, p. 182–197, 2002. Citado nas páginas 91, 92 e 119.

DEVI, A. L.; BOUSSETA, H.; ZATNI, A.; AMGHAR, A.; MOUMEN, A.; ELYAMANI, A.; ZAKARIA, N. A.; HASSAN, W.; HALIN, I.; SIDEK, R. *et al.* Placement and sizing of distributed generators in distributed network based on lric and load growth control. **Journal of Theoretical and Applied Information Technology**, v. 47, n. 1, 2013. Citado na página 28.

DIESTE, O.; PADUA, A. G. Developing search strategies for detecting relevant experiments for systematic reviews. In: **First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)**. [s.n.], 2007. p. 215–224. ISSN 1949-3770. Disponível em: <https://doi.org/10.1109/ESEM.2007.19>. Acesso em: 01 ago. 2018. Citado na página 73.

DING, E.-J.; WU, L.-X. *et al.* Research on key technologies of water inrush perception based on mine iot. **Journal of China Coal Society**, Editorial Office of Journal of China Coal Society, v. 38, n. 8, p. 1397–1403, 2013. Citado na página 28.

DOMINGUES, A.; SILVA, F.; SANTOS, L.; SOUZA, R.; COIMBRA, G.; LOUREIRO, A. A. F. Dados geoespaciais: Conceitos e técnicas para coleta, armazenamento, tratamento e visualização. **Sociedade Brasileira de Computação**, 2020. Citado nas páginas 29 e 32.

DUAN, C.; ZHANG, J.; CHEN, Y.; LANG, Q.; ZHANG, Y.; WU, C.; ZHANG, Z. Comprehensive risk assessment of urban waterlogging disaster based on mcda-gis integration: The case study of changchun, china. **Remote Sensing**, MDPI, v. 14, n. 13, p. 3101, 2022. Citado na página 72.

EASTMAN, J. R.; JIANG, H.; TOLEDANO, J. Multi-criteria and multi-objective decision making for land allocation using gis. **Multicriteria analysis for land-use management**, Springer, p. 227–251, 1998. Citado na página 58.

EIBEN, A. E.; SMITH, J. E. **Introduction to evolutionary computing**. [S.l.]: Springer, 2015. Citado nas páginas 59, 65, 66, 68 e 69.

ESHELMAN, L. J.; SCHAFFER, J. D. Real-coded genetic algorithms and interval-schemata. In: **Foundations of genetic algorithms**. [S.l.]: Elsevier, 1993. v. 2, p. 187–202. Citado na página 67.

ESKELINEN, P.; MIETTINEN, K. Trade-off analysis approach for interactive nonlinear multiobjective optimization. **OR spectrum**, Springer, v. 34, n. 4, p. 803–816, 2012. Citado na página 120.

FEDERAL, D. Portaria n^{o} 2.488, de 14 de outubro de 2011. **Aprova a Política Nacional de Atenção Básica, estabelecendo a revisão de diretrizes e normas para a organização da Atenção Básica, para a Estratégia Saúde da Família (ESF) e o Programa de Agentes Comunit´rio de Sa˜ (PACS). Diário Oficial do Distrito Federal, Brasília, DF**, 2011. Citado na página 103.

FELONI, E. G.; KARPOUZOS, D. K.; BALTAS, E. A. Optimal hydrometeorological station network design using gis techniques and multicriteria decision analysis. **Journal of Hazardous, Toxic, and Radioactive Waste**, American Society of Civil Engineers, v. 22, n. 3, p. 04018007, 2018. Citado nas páginas 78 e 81.

FERREIRA, A.; CHACHADI, A. Assessing aquifer vulnerability to sea-water intrusion using galdit method: Part 2–galdit indicators description. In: **Proceedings of the 4th inter Celtic colloquium on hydrology and management of water resources**. [S.l.: s.n.], 2005. Citado na página 82.

FISCHER, M. M.; WANG, J. **Spatial data analysis: models, methods and techniques**. [S.l.]: Springer Science & Business Media, 2011. Citado nas páginas 34, 35, 36, 37, 38, 43, 44, 45, 46 e 47.

FORMAN, E. H.; GASS, S. I. The analytic hierarchy process—an exposition. **Operations research**, Informs, v. 49, n. 4, p. 469–486, 2001. Citado na página 55.

FOTHERINGHAM, A. S.; BRUNSDON, C.; CHARLTON, M. **Quantitative geography: perspectives on spatial data analysis**. [S.l.]: Sage, 2000. Citado nas páginas 50 e 51.

FREEDMAN, D.; DIACONIS, P. On the histogram as a density estimator: L 2 theory. **Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete**, Citeseer, v. 57, n. 4, p. 453–476, 1981. Citado na página 37.

GASPAR-CUNHA, A.; TAKAHASHI, R.; ANTUNES, C. H. **Manual de computação evolutiva e metaheurística**. [S.l.]: Imprensa da Universidade de Coimbra/Coimbra University Press, 2012. Citado nas páginas 30, 60, 61, 62, 64, 65, 67, 68, 69, 70 e 91.

GEARY, R. C. The contiguity ratio and statistical mapping. **The incorporated statistician**, JSTOR, v. 5, n. 3, p. 115–146, 1954. Citado nas páginas 29 e 46.

GETIS, A.; ORD, J. K. The analysis of spatial association by use of distance statistics. In: **Perspectives on spatial data analysis**. [S.l.]: Springer, 2010. p. 127–145. Citado na página 49.

GIGERENZER, G.; GAISSMAIER, W. Heuristic decision making. **Annual review of psychology**, Annual Reviews, v. 62, p. 451–482, 2011. Citado na página 60.

GOLBERG, D. E. Genetic algorithms in search, optimization, and machine learning. **Addion wesley**, v. 1989, n. 102, p. 36, 1989. Citado nas páginas 61, 65 e 66.

GOLDBARG, E.; GOLDBARG, M.; LUNA, H. **Otimização combinatória e metaheurísticas: algoritmos e apliacações**. [S.l.]: Elsevier Brasil, 2017. Citado nas páginas 60, 61, 62, 63, 64, 65 e 67.

GOLDBERG, D. E. **The design of innovation: Lessons from and for competent genetic algorithms**. [S.l.]: Springer, 2002. v. 1. Citado na página 61.

GÜNEN, M. A. Evaluation of gis based ranking and ahp methods in selecting the most suitable site: A case study in kayseri, turkey. 2021. Citado nas páginas 79 e 84.

HARKER, P. T. The art and science of decision making: The analytic hierarchy process. **The analytic hierarchy process: applications and studies**, Springer, p. 3–36, 1989. Citado na página 55.

HAUPT, R. L.; HAUPT, S. E. **Practical genetic algorithms**. [S.l.]: John Wiley & Sons, 2004. Citado na página 93.

HE, H.; LIU, T.; DU, P. Research on the construction of knowledge graph based on multi-source heterogeneous geospatial data. In: SPIE. **International Conference on Remote Sensing, Surveying, and Mapping (RSSM 2023)**. [S.l.], 2023. v. 12710, p. 112–117. Citado na página 32.

HE, R.; XU, Y.; JIANG, S. Applications of gis in public security agencies in china. **Asian Journal of Criminology**, Springer, v. 17, n. 2, p. 213–235, 2022. Citado na página 28.

HERNANDES, E.; ZAMBONI, A.; FABBRI, S.; THOMMAZO, A. D. Using gqm and tam to evaluate start - a tool that supports systematic review. **CLEI Electronic Journal**, Centro Latinoamericano de Estudios en Informática, v. 15, n. 1, p. 3–3, 2012. Disponível em: <http://www.scielo.edu.uy/scielo.php?script=sci_arttext&pid=S0717-50002012000100003&nrm=iso>. Acesso em: 01 ago. 2018. Citado na página 74.

HO, W.; XU, X.; DEY, P. K. Multi-criteria decision making approaches for supplier evaluation and selection: A literature review. **European Journal of operational research**, Elsevier, v. 202, n. 1, p. 16–24, 2010. Citado na página 69.

HOLLAND, J. H. Robust algorithms for adaptation set in a general formal framework. In: IEEE. **1970 IEEE Symposium on Adaptive Processes (9th) Decision and Control**. [S.l.], 1970. p. 175–175. Citado na página 61.

HONGOH, V.; HOEN, A. G.; AENISHAENSLIN, C.; WAAUB, J.-P.; BÉLANGER, D.; MICHEL, P. Spatially explicit multi-criteria decision analysis for managing vector-borne diseases. **International Journal of Health Geographics**, BioMed Central, v. 10, n. 1, p. 1–9, 2011. Citado na página 72.

HUANG, J.; NIU, L.; ZHAN, J.; PENG, X.; BAI, J.; CHENG, S. Technical aspects and case study of big data based condition monitoring of power apparatuses. In: IEEE. **2014 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)**. [S.l.], 2014. p. 1–4. Citado na página 28.

IPCC. **Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change**. [S.l.], 2021. 1300 p. Citado na página 126.

JANKOWSKI, P.; RICHARD, L. Integration of gis-based suitability analysis and multicriteria evaluation in a spatial decision support system for route selection. **Environment and Planning B: Planning and Design**, SAGE Publications Sage UK: London, England, v. 21, n. 3, p. 323–340, 1994. Citado na página 58.

JENKS, G. F.; CASPALL, F. C. Error on choroplethic maps: definition, measurement, reduction. **Annals of the Association of American Geographers**, Taylor & Francis, v. 61, n. 2, p. 217–244, 1971. Citado na página 41.

JIANG, J.; SHENG, B.; YANG, M. Research on owa based multi-source heterogeneous data fusion. **Advances in Systems Science and Applications**, v. 11, n. 3-4, p. 232–239, 2011. Citado na página 28.

JONG, K. A. D. **An analysis of the behavior of a class of genetic adaptive systems.** [S.l.]: University of Michigan, 1975. Citado na página 68.

KABAK, M.; ERBAŞ, M.; CETINKAYA, C.; ÖZCEYLAN, E. A gis-based mcdm approach for the evaluation of bike-share stations. **Journal of cleaner production**, Elsevier, v. 201, p. 49–60, 2018. Citado nas páginas 79 e 84.

KIM, W.; IIZUMI, T.; NISHIMORI, M. Global patterns of crop production losses associated with droughts from 1983 to 2009. **Journal of Applied Meteorology and Climatology**, American Meteorological Society, Boston MA, USA, v. 58, n. 6, p. 1233 – 1244, 2019. Disponível em: <https://journals.ametsoc.org/view/journals/apme/58/6/jamc-d-18-0174.1.xml>. Citado na página 126.

KITCHENHAM, B.; BRERETON, O. P.; BUDGEN, D.; TURNER, M.; BAILEY, J.; LINKMAN, S. Systematic literature reviews in software engineering – a systematic literature review. **Information and Software Technology**, Elsevier B.V., v. 51, n. 1, p. 7–15, 2009. ISSN 0950-5849. Special Section - Most Cited Articles in 2002 and Regular Research Papers. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950584908001390>. Acesso em: 01 ago. 2018. Citado nas páginas 72 e 73.

KITCHENHAM, B.; CHARTERS, S. **Guidelines for performing Systematic Literature Reviews in Software Engineering**. [S.l.], 2007. Citado na página 72.

KULLDORFF, M. A spatial scan statistic. **Communications in Statistics-Theory and methods**, Taylor & Francis, v. 26, n. 6, p. 1481–1496, 1997. Citado nas páginas 53 e 54.

KULLDORFF, M.; NAGARWALLA, N. Spatial disease clusters: detection and inference. **Statistics in medicine**, Wiley Online Library, v. 14, n. 8, p. 799–810, 1995. Citado nas páginas 53 e 137.

KULLDORFF, M.; RAND, K.; WILLIAMS, G. Satscan: software for the spatial and space-time scan statistics. **Silver Spring, MD: Information Management Services Inc**, 2006. Citado na página 53.

LI, D. A review of high resolution optical satellite surveying and mapping technology. **Spacecraft Recovery & Remote Sensing**, v. 41, n. 2, p. 1–11, 2020. Citado na página 32.

LI, W.-C.; PAN, Y.-c. Multi-source heterogeneous data fusion model in mobile geographic information system. **Journal of Computer Applications**, v. 32, n. 09, p. 2672, 2012. Citado na página 28.

LIMA, L. M. M. d.; MELO, A. C. O. d.; VIANNA, R. P. d. T.; MORAES, R. M. d. Análise espacial das anomalias congênitas do sistema nervoso. **Cadernos Saúde Coletiva**, SciELO Brasil, v. 27, p. 257–263, 2019. Citado nas páginas 101 e 102.

LIMA, R. P. **O processo e o (des) controle da expansão urbana de São Carlos (1857-1977)**. Tese (Doutorado) — Universidade de São Paulo, 2007. Citado na página 140.

LIU, W.; ZHU, J. A multistage decision-making method for multi-source information with shapley optimization based on normal cloud models. **Applied Soft Computing**, Elsevier, v. 111, p. 107716, 2021. Citado nas páginas 87 e 101.

LOBO-FERREIRA, J. P.; CHACHADI, A.; DIAMANTINO, C.; HENRIQUES, M. Assessing aquifer vulnerability to seawater intrusion using galdit method. part 1: application to the portuguese aquifer of monte gordo. 2005. Citado na página 82.

LONGLEY, P. A.; GOODCHILD, M. F.; MAGUIRE, D. J.; RHIND, D. W. **Geographic information systems and science**. [S.l.]: John Wiley & Sons, 2005. Citado nas páginas 37, 38, 39, 40, 41, 42, 43 e 44.

____. **Geographic information science and systems**. [S.l.]: John Wiley & Sons, 2015. Citado na página 90.

LOPES, G. R.; PELARIGO, K. J.; DELBEM, A. C.; SOUSA, J. B. de. Análise exploratória de dados espaciais com python. **Sociedade Brasileira de Computação**, 2022. Citado nas páginas 37, 38, 40, 42, 43, 44 e 122.

LOPES, G. R.; SILVA, R. F. da; PELARIGO, K. J.; YAMAMURA, M.; DELBEM, A. C.; SARAIVA, A. M. Identification of risk areas using spatial clustering to improve dengue monitoring in urban environments. **Revista de Sistemas e Computação-RSC**, v. 12, n. 3, 2023. Citado na página 54.

LOPES, G. R.; SILVA, R. F. da; PELARIGO, K. J.; YAMAMURA, M.; DELBEM, A. C.; SCATOLINI, D.; GHIGLIENO, F.; SARAIVA, A. M. Proposal of a framework for improving multi-criteria decision-making related to epidemics using heterogeneous spatial data and evolutionary algorithms. **Research, Society and Development**, v. 12, n. 2, p. e0212239844–e0212239844, 2023. Citado na página 47.

LU, L.-z.; ZHANG, H.-K.; SHA, X.-L.; WU, H.-H. Design and implementation of a comprehensive information system for detection of urban active faults. In: IEEE. **2012 International Conference on Biomedical Engineering and Biotechnology**. [S.l.], 2012. p. 1311–1314. Citado na página 28.

LUCENA, S. E. d. F.; MORAES, R. M. d. Detecção de agrupamentos espaço-temporais para identificação de áreas de risco de homicídios por arma branca em joão pessoa, pb. **Boletim de Ciências Geodésicas**, SciELO Brasil, v. 18, p. 605–623, 2012. Citado na página 137.

MALCZEWSKI, J. Gis-based multicriteria decision analysis: a survey of the literature. **International journal of geographical information science**, Taylor & Francis, v. 20, n. 7, p. 703–726, 2006. Citado nas páginas 58 e 63.

MALCZEWSKI, J.; RINNER, C. **Multicriteria decision analysis in geographic information science**. [S.l.]: Springer, 2015. v. 1. Citado nas páginas 29, 54, 55, 56, 58, 59, 60, 67, 69, 71, 72, 87, 91 e 118.

MARK, D. M. Geographic information science: Defining the field. **Foundations of geographic information science**, Taylor and Francis, New York, v. 1, p. 3–18, 2003. Citado nas páginas 54 e 55.

MARQUES-PEREZ, I.; GUAITA-PRADAS, I.; GALLEGO, A.; SEGURA, B. Territorial planning for photovoltaic power plants using an outranking approach and gis. **Journal of Cleaner Production**, Elsevier, v. 257, p. 120602, 2020. Citado nas páginas 79 e 83.

MATSUMOTO, P. S. S.; CATÃO, R. da C.; GUIMARÃES, R. B. Mentiras com mapas na geografia da saúde: métodos de classificação e o caso da base de dados de lva do sinan e do cve. **Hygeia: Revista Brasileira de Geografia Medica e da Saude**, Associação Nacional de Pesquisa e Pós-Graduação em Geografia, Grupo de . . . , v. 13, n. 26, p. 211, 2017. Citado nas páginas 105, 116 e 130.

MELO, A. C. O. d.; MELO, J. C. d. S.; MORAES, R. Epidemiologia espacial e a detecção de aglomerados espaciais do dengue na paraíba: uma comparação entre os métodos scan flexível e scan circular. **Cadernos Saúde Coletiva**, SciELO Brasil, v. 30, p. 561–571, 2022. Citado nas páginas 101 e 102.

MESSAOUDI, D.; SETTOU, N.; NEGROU, B.; SETTOU, B. Gis based multi-criteria decision making for solar hydrogen production sites selection in algeria. **International Journal of Hydrogen Energy**, Elsevier, v. 44, n. 60, p. 31808–31831, 2019. Citado nas páginas 79 e 83.

MILOSLAVSKAYA, N.; TOLSTOY, A. Big data, fast data and data lake concepts. **Procedia Computer Science**, Elsevier, v. 88, p. 300–305, 2016. Citado na página 94.

MITCHELL, M. **An introduction to genetic algorithms**. [S.l.]: MIT press, 1998. Citado nas páginas 61, 63 e 64.

MOGHADDAM, H. K.; JAFARI, F.; JAVADI, S. Vulnerability evaluation of a coastal aquifer via galdit model and comparison with drastic index using quality parameters. **Hydrological Sciences Journal**, Taylor & Francis, v. 62, n. 1, p. 137–146, 2017. Citado na página 82.

MONTEIRO, A. M. V.; CÂMARA, G.; CARVALHO, M.; DRUCK, S. Análise espacial de dados geográficos. **Brasília: Embrapa**, 2004. Citado nas páginas 29, 32 e 33.

MORAES, R. M.; NOGUEIRA, J. A.; SOUSA, A. C. A new architecture for a spatio-temporal decision support system for epidemiological purposes. In: WORLD SCIENTIFIC. **Decision Making and Soft Computing: Proceedings of the 11th International FLINS Conference**. [S.l.], 2014. p. 17–23. Citado na página 87.

MORAN, P. A. The interpretation of statistical maps. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, v. 10, n. 2, p. 243–251, 1948. Citado nas páginas 29, 45 e 47.

MU, E.; PEREYRA-ROJAS, M.; MU, E.; PEREYRA-ROJAS, M. Understanding the analytic hierarchy process. **Practical decision making: an introduction to the analytic hierarchy process (AHP) using super decisions V2**, Springer, p. 7–22, 2017. Citado na página 56.

MUNIER, N.; HONTORIA, E. *et al.* **Uses and Limitations of the AHP Method**. [S.l.]: Springer, 2021. Citado na página 119.

NEDELJKOVIC, V.; MILOSAVLJEVIC, M. On the influence of the training set data preprocessing on neural networks training. In: IEEE COMPUTER SOCIETY. **11th IAPR International Conference on Pattern Recognition. Vol. II. Conference B: Pattern Recognition Methodology and Systems**. [S.l.], 1992. v. 1, p. 33–34. Citado na página 28.

NYIMBILI, P. H.; ERDEN, T. Gis-based fuzzy multi-criteria approach for optimal site selection of fire stations in istanbul, turkey. **Socio-Economic Planning Sciences**, Elsevier, v. 71, p. 100860, 2020. Citado nas páginas 78 e 82.

ODEH, T.; SAWAQED, R.; MURSHID, E. A.; MOHAMMAD, A. H. Gis-based analytical analysis for selecting potential runoff harvesting sites: the case study of amman-zarqa basin. **Sustainable Water Resources Management**, Springer, v. 9, n. 3, p. 1–15, 2023. Citado nas páginas 78 e 81.

ODU, G. Weighting methods for multi-criteria decision making technique. **Journal of Applied Sciences and Environmental Management**, v. 23, n. 8, p. 1449–1457, 2019. Citado na página 118.

OPENSHAW, S.; CHARLTON, M.; WYMER, C.; CRAFT, A. A mark 1 geographical analysis machine for the automated analysis of point data sets. **International Journal of Geographical Information System**, Taylor & Francis, v. 1, n. 4, p. 335–358, 1987. Citado na página 53.

O'SULLIVAN, D.; UNWIN, D. **Geographic information analysis**. [S.l.]: John Wiley & Sons, 2003. Citado na página 90.

PANAHI, M.; YEKRANGNIA, M.; BAGHERI, Z.; POURGHASEMI, H. R.; REZAIE, F.; AGHDAM, I. N.; DAMAVANDI, A. A. Gis-based swara and its ensemble by rbf and ica data-mining techniques for determining suitability of existing schools and site selection of new school buildings. In: **Spatial Modeling in GIS and R for Earth and Environmental Sciences**. [S.l.]: Elsevier, 2019. p. 161–188. Citado nas páginas 78 e 81.

PARETO, V. **Cours d'économie politique**. [S.l.]: Librairie Droz, 1964. v. 1. Citado na página 69.

PETERSEN, K.; FELDT, R.; MUJTABA, S.; MATTSSON, M. Systematic mapping studies in software engineering. In: **Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering**. Swindon, UK: BCS Learning & Development Ltd., 2008. (EASE'08), p. 68–77. Disponível em: <http://dl.acm.org/citation.cfm?id=2227115.2227123>. Acesso em: 01 ago. 2018. Citado nas páginas 72 e 75.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and Software Technology**, v. 64, p. 1 – 18, 2015. ISSN 0950-5849. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950584915000646>. Acesso em: 01 ago. 2018. Citado nas páginas 72 e 73.

PFEIFFER, D. U.; ROBINSON, T. P.; STEVENSON, M.; STEVENS, K. B.; ROGERS, D. J.; CLEMENTS, A. C. **Spatial analysis in epidemiology**. [S.l.]: OUP Oxford, 2008. Citado nas páginas 34, 53 e 54.

RAHMAN, A. ur. Geo-spatial disease clustering for public health decision making. **Informatica**, v. 46, n. 6, 2022. Citado na página 72.

RAUF, L. F.; ALI, S. S.; AL-ANSARI, N. Indicator of suitability for evaluating the aquifer thermal energy storage using the gis-based mcda technique in the halabja-khurmal sub-basin. **Applied Water Science**, Springer, v. 13, n. 5, p. 1–10, 2023. Citado nas páginas 80 e 85.

SAATY, R. W. The analytic hierarchy process—what it is and how it is used. **Mathematical modelling**, Elsevier, v. 9, n. 3-5, p. 161–176, 1987. Citado nas páginas 55, 56, 57 e 58.

SAATY, T. L. A scaling method for priorities in hierarchical structures. **Journal of mathematical psychology**, Elsevier, v. 15, n. 3, p. 234–281, 1977. Citado nas páginas 113 e 118.

____. Modeling unstructured decision problems—the theory of analytical hierarchies. **Mathematics and computers in simulation**, Elsevier, v. 20, n. 3, p. 147–158, 1978. Citado nas páginas 113 e 128.

____. **What is the analytic hierarchy process?** [S.l.]: Springer, 1988. Citado nas páginas 55, 56 e 58.

____. How to make a decision: the analytic hierarchy process. **European journal of operational research**, Elsevier, v. 48, n. 1, p. 9–26, 1990. Citado nas páginas 55, 56 e 58.

____. Decision making with the analytic hierarchy process. **International journal of services sciences**, Inderscience Publishers, v. 1, n. 1, p. 83–98, 2008. Citado na página 56.

SAATY, T. L. *et al.* **Decision making with dependence and feedback: The analytic network process**. [S.l.]: RWS publications Pittsburgh, 1996. v. 4922. Citado na página 81.

SAATY, T. L.; VERGAS, L. Multicriteria decision making: the analytical hierarchical process. **RWS, Pittsburg. 125p**, 1992. Citado na página 113.

SCHENK, L.; FANTIN, M.; PERES, R. A revisão do plano diretor da cidade de são carlos e as novas formas urbanas em curso. **Anais do X Colóquio Quapá-SEL: produção e apropriação dos espaços livres e da forma urbana**, Fauunb Brasília, 2015. Citado na página 140.

SEADE. Fundação sistema estadual de análise de dados. Índice paulista de vulnerabilidade social. 2010. Citado nas páginas 16, 107 e 111.

SEIFERT-DÄHNN, I. Insurance engagement in flood risk reduction – examples from household and business insurance in developed countries. **Natural Hazards and Earth System Sciences**, v. 18, n. 9, p. 2409–2429, 2018. Disponível em: <https://www.nat-hazards-earth-syst-sci.net/18/2409/2018/>. Citado na página 126.

SHISHIDO, H. Y. **Escalonamento de workflow com anotações de tarefas sensitivas para otimização de segurança e custo em nuvens**. Tese (Doutorado) — Universidade de São Paulo, 2018. Citado na página 67.

SILVA, R. F.; BENSO, M. R.; GESUALDO, G. C.; MENDIONDO, E. M.; SARAIVA, A. M.; MARQUES, P. A.; DELBEM, A. C. Multi-objective methods for crop insurance premiums: framework proposal and a case study in sugarcane. In: SBC. **Anais do XIII Congresso Brasileiro de Agroinformática**. [S.l.], 2021. p. 225–233. Citado na página 127.

SOHA, T.; HARTMANN, B. Complex power-to-gas plant site selection by multi-criteria decision-making and gis. **Energy Conversion and Management: X**, Elsevier, v. 13, p. 100168, 2022. Citado nas páginas 79 e 84.

TALBI, E.-G. **Metaheuristics: from design to implementation**. [S.l.]: John Wiley & Sons, 2009. Citado na página 63.

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. Citado na página 33.

TORRES-PRECIADO, V. H.; POLANCO-GAYTAN, M.; TINOCO-ZERMEÑO, M. Á. Technological innovation and regional economic growth in mexico: a spatial perspective. **The Annals of Regional Science**, Springer, v. 52, n. 1, p. 183–200, 2014. Citado na página 52.

TURNBULL, B. W.; IWANO, E. J.; BURNETT, W. S.; HOWE, H. L.; CLARK, L. C. Monitoring for clusters of disease: application to leukemia incidence in upstate new york. **American Journal of Epidemiology**, Oxford University Press, v. 132, n. supp1, p. 136–143, 1990. Citado na página 53.

UYDURAN, H. G.; İŞERI, O. K.; ÜSTÜNES, Y.; DURSUN, O. Optimizing wall insulation material parameters in renovation projects using nsga-ii. In: IEEE. **2016 IEEE Congress on Evolutionary Computation (CEC)**. [S.l.], 2016. p. 4208–4213. Citado na página 91.

VANOLYA, N. M.; JELOKHANI-NIARAKI, M.; TOOMANIAN, A. Validation of spatial multi-criteria decision analysis results using public participation gis. **Applied Geography**, Elsevier, v. 112, p. 102061, 2019. Citado na página 72.

VARGAS, L. G. An overview of the analytic hierarchy process and its applications. **European journal of operational research**, Elsevier, v. 48, n. 1, p. 2–8, 1990. Citado na página 59.

WHITLEY, L. D. *et al.* **The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best.** [S.l.]: Colorado State University, Department of Computer Science, 1989. Citado na página 68.

WU, S.-J.; CHOW, P.-T. Steady-state genetic algorithms for discrete optimization of trusses. **Computers & structures**, Elsevier, v. 56, n. 6, p. 979–991, 1995. Citado na página 68.

XIAO-BIN, X.; CHENG-LIN, W. *et al.* A data fusion algorithm of heterogeneous multisensor system based on uniform description of multisource information. **Journal of Henan University (Natural Science)**, 2005. Citado na página 32.

YALEW, S.; GRIENSVEN, A. V.; ZAAG, P. van der. Agrisuit: A web-based gis-mcda framework for agricultural land suitability assessment. **Computers and Electronics in Agriculture**, Elsevier, v. 128, p. 1–8, 2016. Citado na página 72.

ZBIGNIEW, M. Genetic algorithms+ data structures= evolution programs. **Comput Stat**, p. 372–373, 1996. Citado na página 64.

ZHANG, B.; ZHANG, Q.-Q.; CAI, Y.-Y.; YAN, X.-T.; ZHAI, Y.-Q.; GUO, Z.; YING, G.-G. Environmental emissions and pollution characteristics of mosquitocides for the control of dengue fever in a typical urban area. **Science of The Total Environment**, Elsevier, p. 161513, 2023. Citado na página 114.

ZHANG, J. Multi-source remote sensing data fusion: status and trends. **International Journal of Image and Data Fusion**, Taylor & Francis, v. 1, n. 1, p. 5–24, 2010. Citado nas páginas 31 e 32.

ZHANG, L.; XIE, Y.; XIDAO, L.; ZHANG, X. Multi-source heterogeneous data fusion. In: **2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)**. [S.l.: s.n.], 2018. p. 47–51. Citado nas páginas 31 e 32.

ZHANG, X.; RAO, H.; WU, Y.; HUANG, Y.; DAI, H. Comparison of spatiotemporal characteristics of the covid-19 and sars outbreaks in mainland china. **BMC infectious diseases**, Springer, v. 20, n. 1, p. 1–7, 2020. Citado na página 52.

ZHANG, Y.-c.; XING, T.-t. A new method on analyzing modeling of multi-source information in complicated system. **ACTA ELECTONICA SINICA**, v. 37, n. 11, p. 2427, 2009. Citado na página 31.

ZHENG, Y.; CAPRA, L.; WOLFSON, O.; YANG, H. Urban computing: concepts, methodologies, and applications. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM New York, NY, USA, v. 5, n. 3, p. 1–55, 2014. Citado na página 32.

ZUBEN, F. J. V. Computação evolutiva: uma abordagem pragmática. **Anais da I Jornada de Estudos em Computação de Piracicaba e Região (1a JECOMP)**, v. 1, p. 25–45, 2000. Citado na página 64.


ICMC
USP
SÃO CARLOS