Análise da distribuição geográfica de doenças causadoras de óbitos na cidade de São Paulo utilizando aprendizado de máquina

Giovana Martinelli

Dissertação apresentada ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para obtenção do título de
Mestra em Ciências

Programa: Ciência da Computação
Orientador: Prof. Fabio Kon
Coorientador: Prof. Raphael Yokoingawa de Camargo

Esta pesquisa foi financiada por FAPESP, CNPq e CAPES.

São Paulo
Abril de 2023

Análise da distribuição geográfica de doenças causadoras de óbitos na cidade de São Paulo utilizando aprendizado de máquina

Giovana Martinelli

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 14 de Abril de 2023.

Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão julgadora:

Prof. Dr. Fabio Kon (orientador) – IME-USP
Prof. Dr. Paulo Hilário Nascimento Saldiva – FM-USP
Prof. Dr. Rudi Rocha – EAESP-FGV

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Agradecimentos

Gostaria de expressar minha profunda gratidão a todos aqueles que me apoiaram durante a realização deste trabalho de mestrado. Agradeço aos meus orientadores, que me forneceram orientação e suporte ao longo desses três anos. Seus conhecimentos foram fundamentais para o desenvolvimento das minhas habilidades de pesquisa e escrita.

Agradeço também à professora Ligia Vizeu Barrozo da Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH-USP) pela contribuição do seu conhecimento em análises espaciais. Ao professor Rudi Rocha da Escola de Administração de Empresas da Fundação Getúlio Vargas (FGV) e também o nosso ponto focal do Instituto de Estudos para Políticas de Saúde (IEPS) pelas suas constribuições com tópicos relacionados à saúde pública e mortalidade. À professora Daniela Testoni Costa Nobre da Escola Paulista de Medicina da Universidade Federal de São Paulo (UNIFESP) pelas contribuições valiosas após a qualificação. Ao Dr. Rubens Kon, médico da Universidade de São Paulo e médico sanitarista, pela assistência na etapa final do trabalho. Ao Dr. Paulo Saldiva, médico patologista e professor da Faculdade de Medicina da Universidade de São Paulo (FM-USP), pelas ricas sugestões durante a defesa deste trabalho. Sem o apoio desses profissionais, não teria sido possível alcançar esse marco importante em minha vida acadêmica.

Aos membros do projeto InterSCity[1], por compartilharem conhecimentos extremamente valiosos que contribuíram diretamente nesta pesquisa.

Gostaria de agradecer a todos os meus amigos e familiares por seu constante apoio e incentivo, em especial meus pais, Tânia Alexandre Martinelli e Jocimar Martinelli, minha irmã, Fernanda Martinelli, e meu marido, Rafael Rocha da Silva, que estiveram ao meu lado durante todo esse tempo. O apoio que recebi deles foi fundamental para superar os momentos difíceis que surgiram ao longo do processo. Este trabalho é dedicado a todos aqueles que me ajudaram a chegar até aqui.

[1] Esta pesquisa é parte do INCT da Internet do Futuro para Cidades Inteligentes, financiado por CNPq (proc. 465446/2014-0), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e FAPESP (procs. 14/50937-1 e 15/24485-9).

Resumo

Giovana Martinelli. **Análise da distribuição geográfica de doenças causadoras de óbitos na cidade de São Paulo utilizando aprendizado de máquina**. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

Novos estudos na área de epidemiologia estão surgindo constantemente, trazendo grandes contribuições no contexto de saúde pública. Alguns estudos existentes analisam a distribuição geográfica de doenças específicas em determinadas regiões, porém poucos estudos investigam a similaridade da distribuição geográfica de diferentes doenças. Essa falta de pesquisas reflete uma falta de conhecimento sobre como criar políticas públicas específicas com base no território para diferentes grupos de doenças. Este trabalho tem como objetivo estudar, através do uso de análises estatísticas espaciais e aprendizado de máquina, quais doenças causadoras de óbitos na cidade de São Paulo possuem distribuições geográficas similares. Nossa contribuição é dupla: definição de uma nova metodologia para identificar de quais grupos de doenças possuem comportamentos semelhantes em relação às suas distribuições geográficas de óbitos, e também prover um estudo de caso para a cidade de São Paulo. As principais conclusões foram que as maiores taxas de mortalidade associadas às neoplasias estão concentradas na região central da cidade, enquanto as maiores taxas de mortalidade associadas às doenças do sistema circulatório estão concentradas na periferia. Isso mostra que, para as regiões com melhores condições socioeconômicas, a população vai a óbito por causas inevitáveis, que não possuem cura, pois recebem bons tratamentos para as demais causas. Pelo contrário, a população em situação mais vulnerável vai a óbito antes, por uma causa para a qual existe cura e prevenção, porém essas regiões possuem uma maior precariedade no acesso a serviços de saúde. Já com relação às doenças endócrinas, nutricionais e metabólicas, uma parcela das regiões central e oeste possuem taxas de mortalidade um pouco mais baixas em relação às demais regiões. Essas são doenças que são prevenidas com boa alimentação e prática de atividades físicas, o que a população de mais alta renda consegue obter com maior facilidade. Porém, para as doenças do sistema nervoso, essas regiões possuem taxas mais altas que para as demais áreas. Trabalhos futuros podem trazer justificativas desse comportamento para apoiar o desenvolvimento de novas políticas de saúde.

Palavras-chave: Análise Espacial. Epidemiologia. Saúde Pública. SUS. Aprendizado de Máquina. Ciência de Dados. Mortalidade.

Abstract

Giovana Martinelli. **Analysis of the geographic distribution of death-causing diseases in the city of São Paulo using machine learning**. Thesis (Master's). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2023.

New studies in the field of epidemiology are constantly emerging, bringing significant contributions to the context of public health. There are studies that look into the geographic distribution of specific diseases in particular regions, but very few studies investigate the similarity of the geographic distribution of different diseases. This lack of research reflects a lack of knowledge on how to create specific territory-based public policies for different disease groups. This work aims to study, through the use of spatial statistical analysis and machine learning, which diseases causing deaths in the city of São Paulo have similar geographic distributions. Our contribution is twofold: defining a new methodology for identifying which groups of diseases have similar behaviors in relation to their geographic distributions of deaths and also providing a case study for the city of São Paulo. The major findings were that higher mortality rates associated with neoplasms are concentrated in the central region of the city, while higher mortality rates associated with circulatory system diseases are concentrated in the periphery. This shows that, in regions with better socio-economic conditions, the population dies from inevitable causes that have no cure, as they receive good care for the other causes. On the other hand, the population in a more vulnerable situation dies earlier from a cause that has a cure and prevention, but these regions have greater precariousness in accessing health services. As for endocrine, nutritional, and metabolic diseases, a portion of the central and western regions have slightly lower mortality rates than other regions. These are diseases that are prevented with good nutrition and physical activity, which the higher-income population can obtain more easily. However, for nervous system diseases, these regions have higher rates than other areas. Future work may provide justifications for this behavior to support the development of new health policies.

Keywords: Spacial Analysis. Epidemiology. Public Health. SUS. Machine Learning. Data Science. Mortality.

Lista de Abreviaturas

CID	Classificação Internacional de Doenças
OMS	Organização Mundial da Saúde
LISA	Indicadores Locais de Associação Espacial (*Local Indicators of Spatial Association*)
IDH-M	Índice de Desenvolvimento Humano Municipal
PIB	Produto Interno Bruto
DATASUS	Departamento de Informática do Sistema Único de Saúde
SUS	Sistema Único de Saúde
SIM	Sistema de Informações sobre Mortalidade
SEADE	Sistema Estadual de Análise de Dados
SMDU	Secretaria Municipal de Desenvolvimento Urbano
SIG	Sistema de Informações Geográficas
IDH	Índice de Desenvolvimento Humano
UBS	Unidade Básica de Saúde

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Com o grande crescimento da quantidade de informações disponíveis nos dias de hoje, com a evolução da sua qualidade e também com o avanço tecnológico, surgem todos os dias novos estudos nos quais a Ciência de Dados nos ajuda a compreender melhor fenômenos relevantes para a sociedade. Esta pesquisa busca usufruir dessas progressões para trazer novas contribuições para a área da epidemiologia.

Para Czeresnia e Ribeiro, 1997, a epidemiologia pode ser estabelecida como a área da ciência que estuda a distribuição de doenças na população humana e o espaço geográfico está totalmente atrelado a esse conceito. A distribuição geográfica pode ser determinada como a delimitação da área de ocorrência de determinado evento. Para Druck *et al.*, 2004, a maior parte dos problemas de distribuição geográfica levam em consideração três tipos de dados:

- Eventos ou padrões pontuais: são eventos que ocorrem em pontos específicos do espaço, como por exemplo a ocorrência de crimes ou doenças. Neste caso, são utilizadas as localizações exatas das ocorrências.
- Superfícies contínuas: são eventos estimados através de um conjunto de amostras de campo, em geral resultados de estudos de recursos naturais e que incluem mapas topográficos, dentre outros tipos.
- Áreas com contagens e taxas agregadas: são eventos agregados em unidades de análise que representam eventos ocorridos em pontos específicos do espaço. Geralmente são dados associados a levantamentos populacionais, como censos. Na maioria dos casos, as agregações são realizadas por razões de confidencialidade e exemplos comuns são setores censitários, municípios e estados. São esses tipos de dados que serão trabalhados nesta pesquisa.

Silva, 2000 aponta que utilizar o espaço geográfico para o estudo da ocorrência e distribuição das doenças é uma prática que se difundiu antes mesmo da existência da epidemiologia como disciplina científica. As primeiras aplicações reais de análise espacial são atribuídas ao estudo de John Snow (Snow, 1999) sobre a epidemia de cólera ocorrida em Londres em 1854.

Pessoa (1978, *apud* Hino *et al.*, 2011) afirma que atribuir a ocorrência de doenças a

poucos fatores, como apenas a presença de um germe, por exemplo, é uma prática bastante equivocada. É importante levar-se em consideração diversos fatores, como a geografia física da região de análise, fatores meteorológicos, e também as geografias humana, social, política e econômica.

Essa percepção de correlacionar as doenças com o espaço geográfico foi construída ao longo de muito tempo. As causas das doenças já foram consideradas algo sobrenatural, uma designação divina, já acreditou-se que toda doença era causada pelo ar e, depois, que toda doença era causada por um agente biológico (seja ele um vírus ou uma bactéria). Mas foi somente no século XX que Maximiliam Sorre trouxe a principal contribuição para a construção desse contexto de geografia das doenças (Mendonça *et al.*, 2014).

Para Barata, 2013 a epidemiologia se relaciona com as políticas públicas de saúde sob duas perspectivas: como as políticas existentes intervêm no perfil epidemiológico de uma região e como estudos epidemiológicos podem participar na criação e implementação dessas políticas. Ainda segundo a autora, as políticas públicas de saúde abrangem três grandes compromissos: "a redução das desigualdades sociais em saúde, a promoção da saúde e a regulação exercida pelo Estado sobre bens e serviços com consequências sobre a saúde".

Sabendo-se que estamos cercados por uma quantidade significativa de doenças, podemos encontrar um grande número de diferentes distribuições geográficas, dificultando assim, a criação de políticas públicas direcionadas para a atuação sobre cada uma delas. A Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde (representada pela sigla CID 10) é publicada pela Organização Mundial de Saúde (OMS) e contempla códigos relativos às doenças, sintomas e causas externas (CID10, 2013). Essa classificação conta com aproximadamente mil categorias distintas de doenças e vinte e dois capítulos, os quais agrupam doenças com características semelhantes de causa. Visto isso, torna-se inviável a criação de políticas públicas específicas para cada doença e, então, surge a necessidade de identificar quais doenças mais se assemelham e que, consequentemente, podem fazer parte de uma mesma política para que o governo possa atuar de forma direcionada e eficiente.

1.1 Objetivo e motivação

Existem hoje diversos estudos que buscam encontrar a distribuição geográfica de uma doença específica, mas não se sabe, dentre todas as doenças, quais se comportam de maneira semelhante. Alguns exemplos são os estudos de Figueiredo *et al.*, 2001, Souza Dias *et al.*, 2005 e Barcellos *et al.*, 2005 sobre casos de leptospirose, hanseníase e dengue, respectivamente, dentre diversos outros, que serão detalhados mais adiante neste trabalho. Sendo assim, esta pesquisa de mestrado tem como objetivo encontrar quais grupos de doenças possuem distribuições geográficas similares, com a motivação de tornar acessível a criação de um pequeno número de políticas públicas que sejam eficientes para um maior número de doenças possível. Para isso, foram analisados os dados de óbitos de uma grande metrópole, a cidade de São Paulo.

Hoje, sabe-se que a maioria dos óbitos estão relacionados com algum fator que pode ser determinado pela geografia, seja esse fator climático, meteorológico, social, político

ou econômico, segundo Pessoa (1978, *apud* Hino *et al.*, 2011). Por exemplo, regiões com um maior percentual de idosos (fator social) possuem maiores taxas de mortalidade de doenças comuns a esse grupo da população, como o Alzheimer. Também sabe-se que regiões com menor rede de atendimento à saúde (fator político) possuem maiores taxas de mortalidade de fatalidades com necessidade de rápido atendimento, como o infarto. Este trabalho diferencia-se do estado da arte pelo fato de analisar todas as causas de morte em conjunto, e não trabalhar com análises isoladas para uma doença específica ou um grupo de doenças específico. Sendo assim, foram definidas duas questões de pesquisa:

1. Quais grupos de doenças causadoras de morte na cidade de São Paulo possuem distribuições geográficas similares?
2. Quais distritos administrativos da cidade de São Paulo possuem comportamentos similares em relação às doenças causadoras de morte?

As duas questões são complementares com o objetivo de se obterem resultados similares e consistentes sob duas perspectivas diferentes. A hipótese principal para ambas as perguntas é que existem doenças com comportamentos similares entre si em relação às suas distribuições geográficas e é possível identificar esses comportamentos através do uso de algoritmos computacionais.

Neste contexto, serão aplicadas técnicas estatísticas e algoritmos de aprendizado de máquina. O aprendizado de máquina é uma subárea da Inteligência Artificial que se baseia no aprendizado através dos dados para a identificação de padrões e tomadas de decisões com o mínimo de intervenção humana possível. Dentro do aprendizado de máquina existem três grandes subdivisões: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. Esses tipos de algoritmos vêm sendo cada vez mais aplicados em todas as áreas de estudo, inclusive em saúde pública e esta pesquisa trabalhará com aprendizado não supervisionado. Para a aplicação do algoritmo, serão utilizados dados de mortalidade geral da cidade de São Paulo segmentados pelo distrito administrativo de residência do indivíduo e por doença causadora do óbito, extraídos através de bases de dados públicas.

1.2 Contribuições

As principais contribuições desta pesquisa estão concentradas em definir uma nova metodologia que possa ser aplicada em trabalhos futuros e também prover um estudo de caso para a cidade de São Paulo. Espera-se que tanto pesquisadores quanto gestores de saúde possam usufruir dos resultados obtidos.

Para pesquisadores, pretende-se contribuir não só com a metodologia final definida, mas também com todo o aprendizado alcançado acerca dos dados e técnicas utilizados. Além disso, é almejado que esta pesquisa evolua, podendo abranger novas regiões de estudo e novas granularidades de dados.

Já para gestores de saúde, a expectativa é que os resultados possam fornecer informações ainda não conhecidas em relação à distribuição das doenças na cidade de São Paulo e, consequentemente, que auxilie na criação de novas políticas que possam melhorar a saúde pública da cidade.

O próximo capítulo irá abordar os conceitos técnicos necessários para o entendimento da metodologia construída. Em seguida, serão apresentados alguns trabalhos relacionados ao tema desta pesquisa, tanto no Brasil quanto no mundo. Então, será detalhada a metodologia definida para responder às duas questões de pesquisa estabelecidas, assim como apresentação dos dados utilizados e análises preliminares. Por fim, serão apresentados tanto os resultados finais obtidos quanto os resultados negativos que serviram de aprendizado para as análises finais e, então, as conclusões serão evidenciadas no último capítulo.

Capítulo 2

Fundamentação teórica

Neste capítulo apresentamos algumas definições teóricas utilizadas na construção da metodologia, definida no Capitulo 4. Durante a preparação dos dados, alguns tratamentos são necessários. Visto isso, a Seção 2.1 aborda dois métodos utilizados para transformação de dados.

Nosso trabalho utilizou técnicas de aprendizado de máquina e estatística espacial como ferramentas para obter respostas das questões de pesquisa definidas. Como mencionado no Capítulo 1, podemos dividir o aprendizado de máquina em três categorias (Ludermir, 2021):

- Aprendizado supervisionado: os dados possuem rótulos, ou seja, existe uma resposta correta para cada exemplo de dados fornecido ao algoritmo (por exemplo, quando queremos classificar se um conjunto de sintomas corresponde a uma doença específica ou não). Dessa forma, o algoritmo "aprende" sob a supervisão de quem provê os rótulos com as respostas corretas, que são utilizados como exemplos no processo de aprendizado.
- Aprendizado não supervisionado: neste caso, os dados não possuem rótulos. O exemplo mais comum desse tipo de aprendizado são algoritmos de agrupamento, onde os exemplos são analisados e agregados em *clusters* por sua similaridade.
- Aprendizado por reforço: o algoritmo não recebe um rótulo, mas sim uma recompensa ou uma punição sobre a hipótese definida de acordo com os dados.

Sendo assim, neste capítulo apresentamos algumas técnicas de aprendizado não supervisionado que foram testadas em nossa pesquisa (Seção 2.2) e também uma maneira de avaliação de agrupamento de dados (Seção 2.3), utilizada para qualificar o agrupamento realizado. Finalmente, a Seção 2.4 apresenta uma técnica de estatística espacial que também foi explorada nesta pesquisa.

2.1 Transformação dos dados

Visto que as diferentes regiões analisadas podem ter padrões distintos na distribuição da população de acordo com sexo e faixa etária, com o intuito de eliminar possíveis

viéses na análise, foi utilizada uma técnica de ajuste dos dados por essas duas variáveis. Esse ajuste visa calcular uma única taxa de mortalidade para toda a população, porém que leva em consideração as possíveis diferenças de acordo com essas características da população.

Uma segunda transformação nos dados foi realizada com o intuito de alterar a escala das taxas de mortalidade calculadas a fim de evitar possíveis *outliers* e possibilitar uma comparação entre as doenças analisadas sem viéses. A padronização de dados é uma prática comum antes da aplicação de algoritmos de aprendizado de máquina.

2.1.1 Ajuste por sexo e faixa etária

A prática de ajuste dos dados por faixa etária é bastante usual quando estão sendo trabalhados dados que possuem diferentes comportamentos de acordo com essa variável. Um artigo publicado pelo *Statistics Canada* (Escritório Nacional de Estatística do Canadá) (Canada, 2017) sugere uma abordagem onde as populações são ajustadas matematicamente para reproduzirem as mesmas proporções de faixa etária. Este trabalho irá incorporar a categoria de sexo no ajuste, seguindo a mesma linha sugerida pelo artigo. Ou seja, ao invés de ajustarmos as proporções apenas de faixas etárias, ajustaremos as proporções da combinação das categorias de sexo e faixa etária.

Adaptando a metodologia do artigo para o contexto deste trabalho, adotamos os passos abaixo para o cálculo da taxa de mortalidade ajustada por sexo e faixa etária.

1. Cálculo da população total de São Paulo;
2. Cálculo da proporção populacional de cada categoria de sexo e faixa etária para a cidade de São Paulo;
3. Para cada doença e distrito administrativo:
   (a) Cálculo da taxa de mortalidade por 100.000 habitantes para cada categoria de sexo e faixa etária;
   (b) Multiplicação da taxa de mortalidade (a) pela proporção populacional (2);
   (c) Soma da taxa multiplicada (b).

Então, o valor obtido em (c) será a taxa de mortalidade ajustada por sexo e faixa etária. Esse método é ilustrado na Figura 2.1 e nada mais é do que uma média ponderada, porém a ponderação é realizada por uma proporção populacional específica e não a do distrito em análise. Na figura, vemos que as taxas de mortalidade são bastante variadas, com maior peso para a população feminina com 40 a 74 anos, que representa cerca de 50% da população acima de 40 anos de São Paulo. Conforme mostra a tabela, a taxa ajustada final, neste caso, é de 32 mortes por 100 mil habitantes. Se o ajuste não fosse realizado, a taxa seria de 41 mortes por 100 mil habitantes.

Dessa forma, com um único valor temos uma indicação geral da taxa de mortes naquele distrito administrativo que independe do seu perfil de sexo e faixa etária. Isso permite a comparação desses valores para distritos com distribuições discrepantes de sexo e faixa etária, o que é muito comum em cidades como, por exemplo, São Paulo.

						a		b
Capítulo CID 10	Distrito Administrativo	Sexo	Faixa Etária	População	Óbitos	Taxa de Mortalidade	Proporção	Taxa x Proporção
I	Água Rasa	Feminino	40 a 74 anos	20.025	2	10	49,9%	5
I	Água Rasa	Feminino	75 anos ou mais	3.866	6	166	5,8%	10
I	Água Rasa	Masculino	40 a 74 anos	15.849	4	22	41,2%	9
I	Água Rasa	Masculino	75 anos ou mais	1.911	5	262	3,1%	8
I	Água Rasa	-	-	41.652	17	41	100,0%	**32** c

Figura 2.1: *Ilustração do ajuste da taxa de mortalidade por sexo e faixa etária.*

2.1.2 Padronização das doenças por distrito administrativo

É natural que uma mesma doença possua diferentes taxas de mortalidade para cada distrito administrativo. Visto isso, a padronização das taxas das doenças por distrito administrativo busca manter os dados em uma mesma escala para todos os distritos, tornando a variável com uma distribuição Normal com média igual a 0 e desvio padrão igual a 1. Essa padronização é feita da seguinte forma:

$$z = \frac{x - \mu}{\sigma}$$

Na fórmula acima, z representa a taxa de mortalidade padronizada para uma determinada doença em um distrito administrativo, x representa a taxa de mortalidade por 100 mil habitantes antes da padronização e μ e σ representam a média e o desvio padrão, respectivamente, da taxa de mortalidade por 100 mil habitantes de todas as doenças desse distrito administrativo. A padronização das doenças por distrito administrativo foi realizada com as taxas de mortalidade ajustadas por sexo e faixa etária, conforme demonstrado anteriormente.

Na prática, quando a taxa de mortalidade padronizada é acima de 0, concluímos que essa doença possui uma taxa de mortalidade acima da média das taxas das doenças nesse distrito. De forma similar, uma taxa negativa indica que essa doença possui uma taxa de mortalidade abaixo da média desse distrito. Assim, a comparação entre doenças dentro de um mesmo distrito administrativo é facilitada, pois todas estarão em uma mesma escala de valores.

2.2 Aprendizado não supervisionado

Para identificar as distribuições geográficas das doenças foram testados alguns algoritmos de aprendizado não supervisionado. As Seções 2.2.1 e 2.2.2 apresentam dois métodos de agrupamento que se baseiam nas distâncias dos dados para agrupá-los, porém com metodologias distintas. Já a Seção 2.2.3 traz uma abordagem que trata dados categóricos, ou seja, dados que não seguem uma escala de valores, mas sim, representam categorias.

2.2.1 Aprendizado hierárquico

O algoritmo de aprendizado hierárquico busca agrupar os dados de acordo com as suas distâncias. Ele pode ser aglomerativo ou divisivo (Oliveira, 2021), onde:

- Aglomerativo: as unidades amostrais iniciam-se cada uma em um aglomerado (*cluster*) distinto e vão sendo agrupadas iterativamente até que todas façam parte de um único *cluster*.
- Divisivo: as unidades amostrais iniciam-se todas como parte de um único *cluster* e vão sendo divididas iterativamente até que cada uma esteja alocada em um *cluster* distinto.

Neste trabalho foi utilizado o algoritmo aglomerativo, pois é o método que já está implementado na biblioteca de aprendizado de máquina *scikit-learn*[1] do Python (linguagem de programação utilizada). O método utilizado pelo algoritmo baseia-se nos seguintes passos:

1. Atribuir cada unidade amostral a um *cluster*;
2. Calcular as distâncias entre os *clusters*;
3. Agrupar os dois *clusters* com a menor distância calculada;
4. Repetir os passos 2 e 3 até que exista um único *cluster*.

Existem alguns diferentes métodos para o cálculo da distância entre os dados, sendo a distância euclidiana a mais utilizada na literatura, que é definida como a menor distância entre dois pontos. O cálculo dessa métrica é dado por:

$$D = \sqrt{\sum_{k=1}^{n}(x_{rik} - x_{sjk})^2}$$

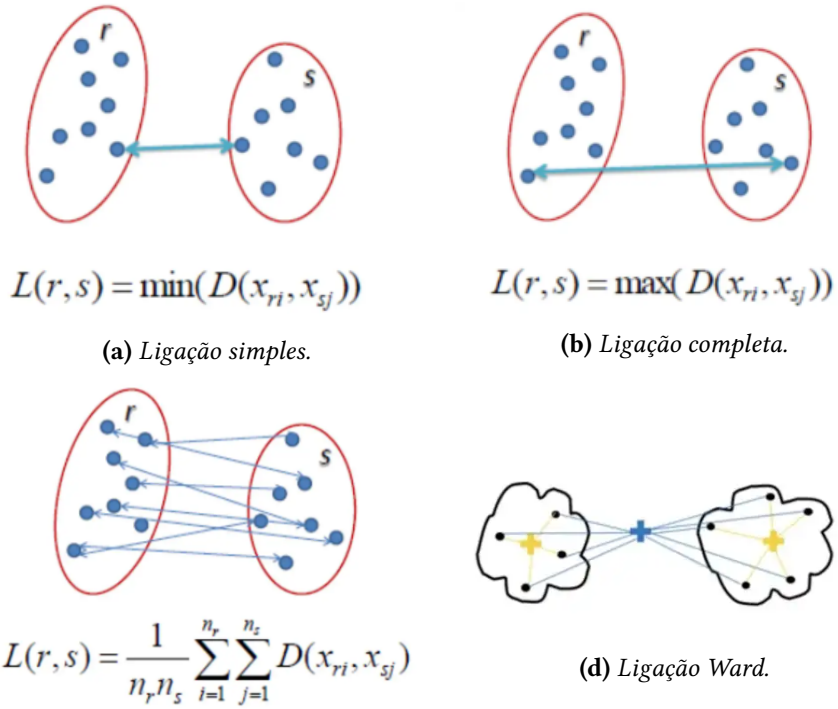
Onde $x_{ri} = \{x_{ri1}, ..., x_{rin}\}$ e $x_{sj} = \{x_{sj1}, ..., x_{sjn}\}$ representam as i-ésima e j-ésima unidades amostrais dos *clusters* r e s, respectivamente.

Também é necessário definir qual será o método de ligação entre os dados (simples, completa, média ou Ward). A ligação entre os *clusters* r e s é representada por $L(r, s)$ e define como os *clusters* serão conectados (Maklin, 2018).

- Simples: a distância entre dois *clusters* é definida pela menor distância entre duas unidades amostrais de cada *cluster* (Figura 2.2a).
- Completa: a distância entre dois *clusters* é definida pela maior distância entre duas unidades amostrais de cada *cluster* (Figura 2.2b).
- Média: a distância entre dois *clusters* é definida pela distância média entre cada unidade amostral de um *cluster* para todas as unidades amostrais do outro *cluster* (Figura 2.2c).

[1] https://scikit-learn.org/

- Ward: a distância entre dois *clusters* é definida pela soma das diferenças ao quadrado entre os dois *clusters* (Figura 2.2d).



(a) *Ligação simples.*

(b) *Ligação completa.*

(c) *Ligação média.*

(d) *Ligação Ward.*

Figura 2.2: *Tipos de ligação no algoritmo de agrupamento hierárquico. Fonte: Maklin, 2018*

O número de *clusters* é definido pelo pesquisador e isso pode ser feito com o auxílio do dendrograma, que é uma ilustração do algoritmo de agrupamento hierárquico. Um exemplo está na Figura 2.3, onde o eixo x do gráfico representa as unidades amostrais e o eixo y representa as distâncias entre as unidades pertencentes aos *clusters* relacionados. Construímos uma ilustração de duas possíveis definições do número de *clusters*. Podemos traçar uma reta em aproximadamente $y = 14$ (reta preta) e definir dois aglomerados e também podemos traçar uma outra reta em aproximadamente $y = 8$ (reta vermelha) e definir quatro aglomerados. Vale ressaltar que não existe um número correto de *clusters*, porém quanto maior o número, menos diferenças teremos dos dados brutos para os dados agrupados e, consequentemente, menor será o ganho dessa segmentação.

2.2.2 K-Means

O K-Means também é um algoritmo que se baseia nas distâncias dos dados, mas que faz isso a partir de centróides. A grande diferença desse método para o agrupamento hierárquico é que, neste caso, o número de *clusters* a serem criados precisa ser definido antes da execução do algoritmo.

O método utilizado pelo algoritmo é chamado particional e segue os passos abaixo, ilustrados também na Figura 2.4 (Piech, 2013).

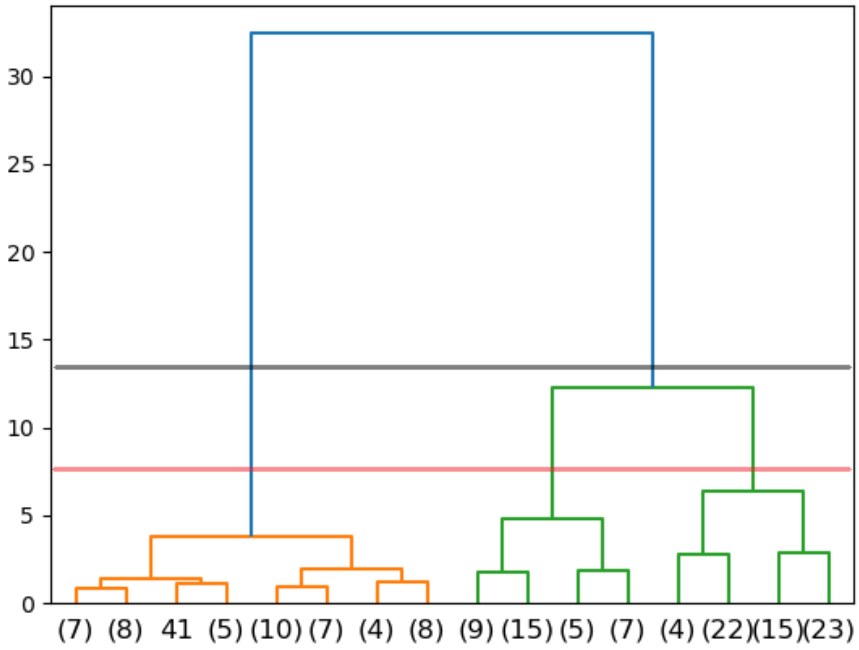


Figura 2.3: *Exemplo de um dendrograma. Fonte: Pedregosa et al., 2011.*

1. As localizações dos centróides dos K *clusters* são geradas aleatoriamente (item (b) da figura para $K = 2$).
2. Os K *clusters* são criados associando cada observação ao centróide mais próximo (item (c) da figura).
3. O centróide de cada *cluster* é alterado para o valor médio dos pontos contidos nele (item (d) da figura).
4. Os passos (2) e (3) são repetidos várias vezes até o algoritmo convergir (itens (e) e (f) da figura).

Os testes realizados com esse algoritmo também utilizam o dendrograma para definir o número ideal de agrupamentos a serem criados.

2.2.3 K-Modes

Esse algoritmo é bastante similar ao K-Means, porém adaptado para lidar com dados categóricos. Neste caso, a métrica utilizada na construção dos *clusters* é a moda ao invés da média. Então, o algoritmo segue os seguintes passos (Aprilliant, 2021):

1. As localizações dos centróides dos K *clusters* são geradas aleatoriamente.
2. Os K *clusters* são criados associando cada observação ao centróide mais próximo (pela métrica de dissimilaridade, apresentada a seguir).
3. O centróide de cada *cluster* é alterado para a moda dos pontos contidos nele.
4. Os passos (2) e (3) são repetidos várias vezes até o algoritmo convergir.

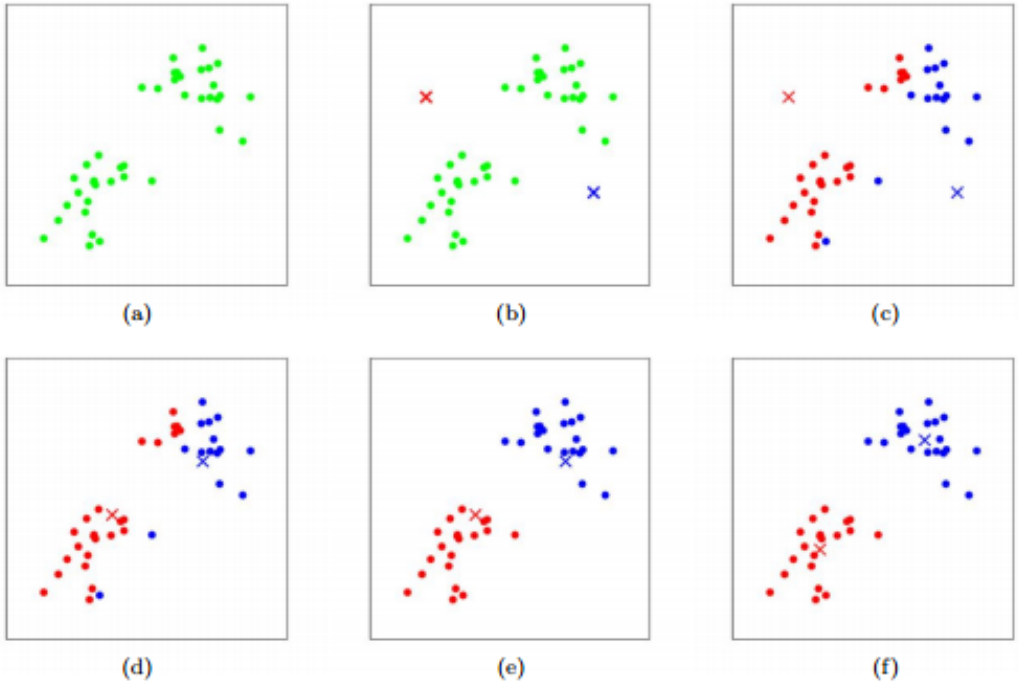


Figura 2.4: *Como funciona o algoritmo K-Means. Fonte: Piech, 2013.*

A métrica de dissimilaridade entre a i-ésima unidade amostral do *cluster r* e a j-ésima unidade amostral do *cluster s* é dada por:

$$D(x_{ri}, x_{sj}) = \sum_{k=1}^{n} \delta(x_{rik}, x_{sjk})$$

Sendo:

$$\delta(x_{rik}, x_{sjk}) = \begin{cases} 0, & \text{se } x_{rik} = x_{sjk} \\ 1, & \text{se } x_{rik} \neq x_{sjk} \end{cases}$$

A Tabela 2.1 ilustra um exemplo da identificação dos centróides dos *clusters* no caso de dados binários.

ID	Categoria 1	Categoria 2	Categoria 3	Categoria 4	Categoria 5
1	1	1	0	1	1
2	1	1	0	0	1
3	1	0	1	0	0
4	1	0	1	0	0
5	1	1	0	0	0
Centróide	1	1	0	0	0

Tabela 2.1: *Exemplo de identificação dos centróides no algoritmo K-Modes. Fonte: Aprilliant, 2021.*

Assim como para o K-Means, os testes realizados com esse algoritmo também utilizam o dendrograma para definir o número ideal de agrupamentos a serem criados.

2.3 Coeficiente de silhueta

Como mencionado anteriormente, não existe um rótulo para os dados no aprendizado não supervisionado e, consequentemente, não faz sentido querermos determinar se o agrupamento criado é o correto. Porém, existem métodos que nos permitem quantificar a similaridade entre os *clusters* criados. Um desses métodos é o coeficiente de silhueta e pode ser aplicado se o número total de agrupamentos criados for entre 2 e o total de amostras menos 1. Portanto, utilizamos esse coeficiente para validar se o agrupamento criado é eficaz.

J.Rousseeuw, 1987 apresenta a forma de cálculo desse coeficiente e, inclusive, é utilizado como referência na documentação da biblioteca *scikit-learn*. Portanto, todas as definições abaixo foram retiradas desse artigo.

Seja i uma unidade amostral disponível no conjunto de dados, $s(i)$ representa o seu coeficiente de silhueta, A corresponde ao *cluster* a qual essa unidade foi alocada e C um outro *cluster* diferente de A. Então, podemos definir:

$a(i)$ = *similaridade média de i para todos os outros objetos em A.*

$d(i, C)$ = *similaridade média de i para todos os objetos em C.*

Na Figura 2.5, $a(i)$ é calculado pela média de todas as linhas que saem de i para os demais pontos em A. Já $d(i, C)$ é calculado pela média de todas as linhas que saem de i para cada ponto em C. Calculados todos os $d(i, C)$ para todos os *clusters* $C \neq A$, é selecionado o maior número, ou seja, a maior similaridade, determinada por $b(i)$.

$$b(i) = \max_{C \neq A} d(i, C)$$

Quanto maior for $d(i, C)$, maior a similaridade da unidade amostral i com o *cluster* C. Isso significa que esse *cluster* seria a segunda opção de alocação para i. Então, o coeficiente de silhueta $s(i)$ é dado por:

$$s(i) = \begin{cases} 1 - b(i)/a(i) & \text{se } a(i) > b(i), \\ 0 & \text{se } a(i) = b(i), \\ a(i)/b(i) - 1 & \text{se } a(i) < b(i). \end{cases}$$

Pode-se ver que $-1 \leq s(i) \leq 1$, onde:

- $s(i)$ positivo indica que o *cluster* mais similar ao alocado para i tem uma similaridade baixa, ou seja, o objeto já está no *cluster* mais apropriado. Quanto mais próximo de 1, maior o indício de que o agrupamento é o mais ideal.
- $s(i)$ igual a 0 indica que não há grandes diferenças entre o *cluster* já alocado e o *cluster* mais similar, ou seja, i poderia pertencer aos dois *clusters*.

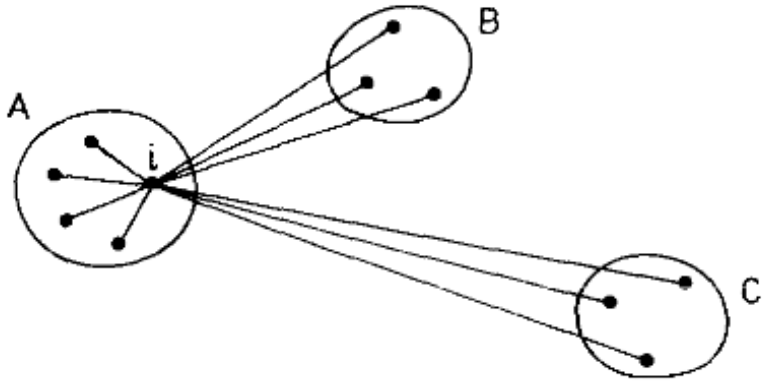


Figura 2.5: *Elementos envolvidos no cálculo de $s(i)$. Fonte: J.Rousseeuw, 1987.*

- $s(i)$ negativo indica que i não está no melhor *cluster*, pois existe um outro ao qual ele é mais similar. Quanto mais próximo de -1, maior o indício de que o agrupamento não é o mais ideal.

Por fim, o coeficiente final utilizado para definir o bom ajuste do agrupamento se dá pela média dos coeficientes $s(i)$ de todas as unidades amostrais. Assim, espera-se obter um coeficiente final próximo de 1 ao final da aplicação do algoritmo de agrupamento.

2.4 Índice de Moran

O Índice de Moran busca adaptar a medida de autocorrelação para o contexto espacial, ou seja, calcula a relação de uma variável com ela mesma, levando em consideração a proximidade espacial das regiões analisadas. Há duas expressões estatísticas distintas para esse índice: o Índice Global, proposto em 1950, e o Índice Local, proposto em 1995 (Luzardo *et al.*, 2017).

As definições teóricas apresentadas a seguir, assim como as fórmulas definidas, foram extraídas do livro *Análise Espacial de Dados Geográficos* de Druck *et al.*, 2004 e do artigo *Análise espacial exploratória com o emprego do índice de moran* de Luzardo *et al.*, 2017.

2.4.1 Matriz de proximidade espacial

O cálculo tanto do Índice Global quanto do Índice Local levam em consideração a matriz de proximidade espacial, também conhecida por matriz de vizinhança. Dado um conjunto de n regiões $\{R_1, ..., R_n\}$, cada elemento w_{ij} da matriz W representa uma medida de proximidade entre as regiões R_i e R_j. Esse relacionamento pode ser representado de diversas formas, os mais comuns são:

- Distância: $w_{ij} = 1$ se o centróide da região R_i estiver dentro de uma distância pré estabelecida do centróide da região R_j. Caso contrário, $w_{ij} = 0$.
- Contiguidade ou adjacência: $w_{ij} = 1$ se a região R_i compartilhar um lado em comum com a região R_j. Caso contrário, $w_{ij} = 0$.

- Perímetro: $w_{ij} = L_{ij}/L_i$, onde L_{ij} é o comprimento da fronteira entre R_i e R_j e L_i é o perímetro de R_i.
- Vizinhança: $w_{ij} = 1$ se a região R_i for um dos k vizinhos mais próximos da região R_j, onde k é um valor pré estabelecido. Caso contrário, $w_{ij} = 0$.

Sendo assim, o primeiro passo antes do cálculo de qualquer um dos índices é definir qual matriz de proximidade espacial será utilizada e construí-la.

2.4.2 Índice Global de Moran

O Índice Global de Moran avalia a autocorrelação espacial da variável de interesse entre todas as regiões da área de estudo, e é expresso por um único valor para toda a área. O índice é calculado através da seguinte fórmula:

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (z_i - \bar{z}) \left(z_j - \bar{z}\right)}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

Onde z_i representa o valor da variável de interesse para a região R_i, $\bar{z}$ representa o valor médio da variável de interesse considerando-se todas as regiões, e w_{ij} representa a proximidade espacial entre as regiões R_i e R_j, como definido na Seção 2.4.1.

Esse índice tem o intuito de avaliar se o comportamento da variável de interesse é aleatório, clusterizado (dados mais agregados que se distribuídos aleatoriamente) ou disperso (dados menos agregados que se distribuídos aleatoriamente). Quando o valor de I é próximo de 0, o comportamento é considerado aleatório, ou seja, existe uma relação de independência espacial da variável analisada em torno das regiões. Quando I é positivo, pode-se dizer que os dados estão clusterizados, ou seja, existe uma autocorrelação direta. Já quando I é negativo, pode-se dizer que os dados estão dispersos, ou seja, existe uma autocorrelação inversa. Através desse índice, conseguiremos saber quais doenças possuem comportamento aleatório e quais podem ser clusterizadas.

2.4.3 Índice Local de Moran

O Índice Local de Moran avalia a autocorrelação espacial entre uma região e as suas vizinhas e, diferentemente do Índice Global de Moran, é expresso por um valor para cada região. O índice para a região i é calculado através da seguinte fórmula:

$$I_i = \frac{z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} z_j^2}$$

Derivado do Índice Local de Moran podemos obter o LISA *Cluster* (*Local Indicators of Spatial Association*), que classifica o índice de cada região em cinco possíveis categorias:

- [AA] Alto-Alto: indica um alto índice para a região de análise e também um alto índice para os seus vizinhos.

- [BB] Baixo-Baixo: indica um baixo índice para a região de análise e também um baixo índice para os seus vizinhos.
- [AB] Alto-Baixo: indica um alto índice para a região de análise, mas um baixo índice para os seus vizinhos.
- [BA] Baixo-Alto: indica um baixo índice para a região de análise, mas um alto índice para os seus vizinhos.
- [NS] Não significativo: indica um índice sem significância estatística para a região de análise.

A Figura 2.6 traz um exemplo construído por Luzardo *et al.*, 2017 de um mapa contendo os LISA *Clusters* para o Índice de Desenvolvimento Humano Municipal (IDH-M) para o estado do Rio de Janeiro. Nele, podemos ver que vários municípios são classificados com índices não significativos (destacados em cinza). Apesar disso, conseguimos observar algumas cidades na região noroeste do estado classificadas como Alto-Alto (em vermelho escuro), ou seja, essa é uma área com altos índices de IDH-M. Também existem alguns municípios isolados com outras classificações, como por exemplo Campos dos Goytacazes (maior município em destaque, na cor vermelho claro) que possui um alto índice de IDH-M, porém suas regiões vizinhas possuem baixos índices.

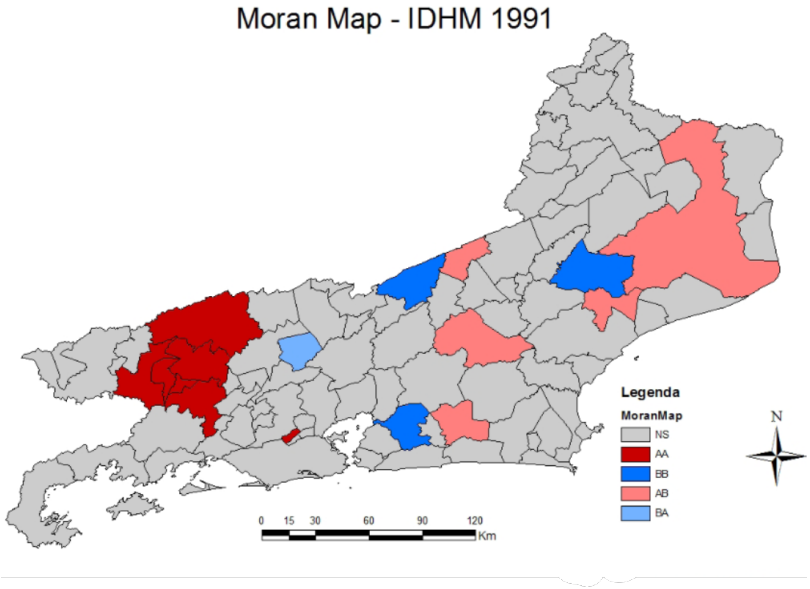


Figura 2.6: *Exemplo de mapa com o LISA Cluster. Fonte:* Luzardo *et al., 2017.*

O cálculo deste índice e do LISA *Cluster* auxiliam na identificação dos *clusters* espaciais, ou seja, quais são as regiões de maior e menor destaque em relação à variável estudada.

O próximo capítulo apresenta os trabalhos relacionados à nossa pesquisa, ou seja, que também abordam a análise geográfica de doenças, tanto no Brasil quanto no mundo, expondo também as técnicas utilizadas para a análise espacial.

Capítulo 3

Trabalhos relacionados

O presente capítulo tem como objetivo apresentar alguns trabalhos desenvolvidos no Brasil e ao redor do mundo, que possuem como temática a análise da distribuição geográfica de doenças. Todas as pesquisas aqui apresentadas abordam a análise de uma doença específica ou mais, existindo uma grande diversidade de doenças estudadas. Ao final do capítulo apresentamos também um estudo que não é na área da epidemiologia, porém tem um objetivo semelhante ao desta pesquisa, que é o agrupamento de regiões com base em suas características.

Um dos primeiros estudos com aplicações reais de análise espacial é o estudo do médico inglês John Snow, sobre a epidemia de cólera em Londres em 1854 (Snow, 1999). Esse foi o ano mais crítico da epidemia e Snow, que já tinha teorias de que a contaminação da doença se dava pela água infectada com fezes, resolveu criar um mapa do bairro de Soho para identificar se alguma região estava sendo mais afetada que outra. Os pontos vermelhos na Figura 3.1 representam os óbitos causados pelo cólera e, em azul, estão representadas as bombas de abastecimento de água. Conforme as mortes iam ocorrendo, percebeu-se que estas estavam concentradas na região central do mapa, com abastecimento de água pela bomba Bread Street. Em uma investigação mais profunda, Snow descobriu que essa bomba estava contaminada pelo esgoto da cidade. Depois de diversas tentativas de provar a sua teoria, esse estudo de John Snow foi reconhecido pela comunidade científica e ele foi considerado o primeiro epidemiologista da história (Almeida, 2018).

Após esse marco histórico, outros estudos surgiram relacionando a ocorrência de doenças com a sua distribuição geográfica, inclusive no Brasil. Figueiredo *et al.*, 2001 buscaram entender a ocorrência geográfica dos casos de leptospirose de 1995 no município de Belo Horizonte, Minas Gerais. As regiões de análise constituíram de nove administrações regionais. Para identificar as regiões de maior foco da doença, foi utilizado um recurso de geoprocessamento chamado cartografia digital e observou-se no mapa o percentual de casos confirmados da doença em relação aos casos suspeitos e, então, foi calculado o intervalo de confiança para cada região. Nesse estudo, é interessante destacar a conclusão de que a ocorrência dessa doença tem maiores ou menores intensidades para diferentes grupos de faixa etária e sexo. Além disso, os resultados apontaram que as regiões sem tratamento de esgoto, com falta de saneamento básico e também as periferias e favelas foram as mais afetadas pela doença. Em termos de saúde pública, esse estudo buscou

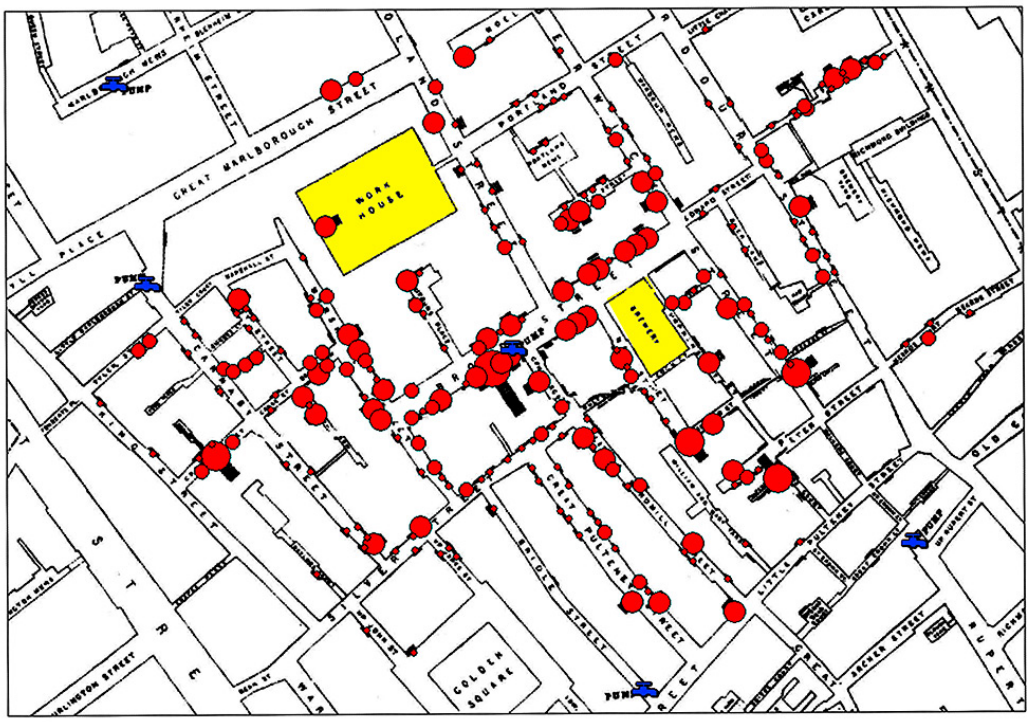


Figura 3.1: *Bairro de Soho criado por John Snow. Fonte:* Almeida, 2018.

o conhecimento das regiões onde a doença ocorreu mais frequentemente para não só subsidiar políticas de saneamento, mas também dar suporte a trabalhos futuros.

Hino *et al.*, 2011 conduziram um estudo sobre a distribuição geográfica dos casos de dengue, hanseníase e tuberculose notificados no ano 2000, no município de Ribeirão Preto, São Paulo. Os endereços exatos dos pacientes foram considerados nesse caso e uma técnica de estatística espacial, chamada de alisamento Kernel, foi utilizada na análise de intensidade das doenças. Essa técnica ajusta uma função bidimensional onde é calculado um valor proporcional de casos em relação à área estudada, realizando uma ponderação pela distância de cada área à uma localização de interesse. Apesar de o foco do estudo não ser em relacionar as distribuições geográficas das três doenças, os resultados mostraram que as regiões de maior incidência do dengue possuem algumas coincidências com as áreas de maior incidência da tuberculose. Já a hanseníase apresentou uma região de destaque diferente das demais. Esse estudo não correlacionou as áreas identificadas com variáveis de sexo e faixa etária, ou indicadores socioeconômicos.

Uma abordagem bastante utilizada para esse tipo de estudo é o uso do Sistema de Informações Geográficas (SIG), sistema que armazena dados georreferenciados e os correlaciona com dados tabulares. Os estudos de Souza Dias *et al.*, 2005, Barcellos *et al.*, 2005 e Jayasekara *et al.*, 2013 são alguns exemplos de aplicação dessa abordagem.

Souza Dias *et al.*, 2005 analisaram a distribuição geográfica da hanseníase entre 1998 e 2002 no município de Mossoró, Rio Grande do Norte. O SIG é utilizado para identificar a distribuição espacial da doença de acordo com o bairro de residência dos pacientes. A análise concluiu que as áreas que apresentaram maiores números de casos da doença correspondem a áreas com baixo padrão socioeconômico e alta densidade demográfica,

e também na região mais endêmica foi identificada uma maior concentração de casos em menores de 15 anos. O trabalho de Barcellos *et al.*, 2005 também produziu análises espaciais sobre os casos de dengue no município de Porto Alegre, Rio Grande do Sul, em 2002, no qual foram utilizados os setores censitários da cidade como regiões de análise. Já o estudo de Jayasekara *et al.*, 2013 foi realizado fora do Brasil, sobre doença renal crônica na região centro-norte do Sri Lanka, país localizado ao sul da Índia. Cinco regiões foram identificadas como destaque e observou-se que aquelas onde a população consumia água proveniente de nascentes naturais tinham menor incidência da doença e que as regiões mais afetadas estavam localizadas abaixo de reservatórios com água estagnada irrigada. Também mostrou-se uma diferença da incidência da doença para diferentes categorias de sexo e faixa etária.

Uma outra abordagem em estatística espacial é o método de krigagem, que usa a função de semivariância para estimar valores da variável de interesse com base nos dados amostrais. A função de semivariância, por sua vez, é utilizada para detectar dependência espacial, ou seja, a relação da variável de interesse com a sua distribuição espacial na região analisada. Os estudos de Opromolla *et al.*, 2006 e Rodrigues-Júnior *et al.*, 2008 utilizam esse método na análise dos casos de hanseníase no estado de São Paulo entre 1991 e 2002 e entre 2004 e 2006, respectivamente.

Opromolla *et al.*, 2006 utiliza os municípios de residência dos pacientes como região de análise e, diferente dos estudos apresentados anteriormente, esse não relaciona as regiões identificadas com as características socioeconômicas dos municípios e também não traz uma perspectiva da análise por sexo e faixa etária. Já o trabalho de Rodrigues-Júnior *et al.*, 2008 agrega os dados em departamentos regionais de saúde para realizar a análise espacial. Neste caso, os resultados mostraram uma correlação positiva entre as regiões com maiores concentrações da doença e variáveis de escolaridade e longevidade, e uma correlação negativa com a variável riqueza, além de identificarem diferenças na ocorrência de hanseníase de acordo com o sexo e faixa etária do paciente.

Diversos estudos utilizam uma outra técnica geoestatística, o Índice de Moran. Como definido no Capítulo 2, esse índice avalia a autocorrelação espacial de uma determinada variável de interesse de acordo com as regiões estudadas. Os estudos de Costa-Nobre *et al.*, 2021, Wang *et al.*, 2012 e KELLY-HOPE *et al.*, 2007 são alguns exemplos de aplicação dessa abordagem.

O estudo de Costa-Nobre *et al.*, 2021 investiga *clusters* de mortalidade neonatal associadas à asfixia no estado de São Paulo entre 2004 e 2013, e explora a sua associação com o produto interno bruto (PIB) per capita. Esse estudo considerou os municípios do estado como granularidade de análise e utilizou autocorrelações espaciais para identificar aqueles com maiores taxas de mortalidade. Também vale mencionar que foi aplicada uma suavização na taxa de mortalidade, através do estimador Bayesiano empírico local, que recalcula as taxas de um município atribuindo um peso às taxas dos municípios vizinhos. Essa taxa suavizada foi capaz de evidenciar ainda mais os clusters identificados e tornou a autocorrelação espacial ainda mais significativa. Também identificou-se uma correlação negativa entre as taxas de mortalidade e o PIB per capita.

Fora do Brasil, Wang *et al.*, 2012 construíram análises sobre a distribuição espacial de tuberculose em Linyi, na China, entre 2005 e 2010. Assim como o trabalho de Costa-Nobre

et al., 2021, foi utilizada a suavização empírica de Bayes no cálculo das taxas e aplicou-se a autocorrelação espacial sobre regiões administrativas da cidade. Através dessa metodologia foi possível identificar cinco *clusters* espaciais, mas nenhuma análise de correlação com indicadores socioeconômicos foi realizada. Já o trabalho de KELLY-HOPE *et al.*, 2007 pesquisou sobre doenças entéricas no Vietnã, país localizado no sudeste asiático, entre 1991 e 2001. A área de estudo é baseada em regiões constituídas pelo agrupamento de províncias do país e a análise de autocorrelação espacial foi aplicada para encontrar distribuições espaciais das doenças analisadas (disenteria, febre tifóide e cólera). Também foi realizada uma análise de correlação com fatores ambientais e humanos. As autocorrelações espaciais foram calculadas para cada uma das doenças separadamente e observou-se que os *clusters* espaciais identificados eram distintos entre as três, possivelmente devido a fatores ecológicos. Para a disenteria, identificou-se que a alta pluviosidade e a pobreza humana são os principais fatores de risco. A febre tifóide teve maior concentração no delta do rio Mekong e foi mais correlacionada com a pressão de vapor e a água potável do rio. Já o cólera se correlacionou positivamente com a chuva e a água potável de poço público.

A Tabela 3.1 traz um breve resumo de todos os trabalhos apresentados anteriormente, listando as doenças abordadas em cada um deles, técnica utilizada para identificar as regiões mais relevantes, áreas de abrangência e períodos abordados.

Doença	Técnica utilizada	Área de abrangência	Período
Leptospirose	Cartografia digital	Município de Belo Horizonte	1995
Dengue, Hanseníase e Tuberculose	Alisamento Kernel	Município de Ribeirão Preto	2000
Hanseníase	SIG	Município de Mossoró	1998 - 2002
Dengue	SIG	Município de Porto Alegre	2002
Doença renal crônica	SIG	Região centro-norte do Sri Lanka	Não informado
Hanseníase	Krigagem	Estado de São Paulo	1991 - 2002
Hanseníase	Krigagem	Estado de São Paulo	2004 - 2006
Asfixia	Índice de Moran	Estado de São Paulo	2004 - 2013
Tuberculose	Índice de Moran	Município de Linyi	2005 - 2010
Disenteria, febre tifóide e cólera	Índice de Moran	Vietnã	1991 - 2001

Tabela 3.1: *Resumo dos trabalhos apresentados.*

Apesar de não ser na área da saúde, o estudo brasileiro de Pena *et al.*, 2017 busca encontrar um agrupamento de áreas mínimas comparáveis (agregado de municípios) com base em variáveis climáticas, de características do solo e produção agropecuária. É importante mencionar esse artigo pois ele traz uma perspectiva multivariada, onde busca o agrupamento de regiões com base em mais de um atributo, diferente dos demais artigos apresentados que sempre buscavam um agrupamento de regiões com base em um único atributo, representado pela taxa de incidência da doença analisada. Nesta pesquisa de

mestrado, também temos um caso multivariado onde os diversos atributos podem ser representados pelas taxas de mortalidade das distintas doenças que estão sob análise. O trabalho de Pena *et al.*, 2017 traz uma comparação entre algoritmos de agrupamento tradicionais e também um algoritmo que incorpora a contiguidade das regiões analisadas, ou seja, busca agregar regiões vizinhas. Sendo assim, são testados três tipos de algoritmos: agrupamento hierárquico não espacial, agrupamento hierárquico espacial (com a incorporação de contiguidade) e k-means (também um algoritmo não supervisionado de agrupamento, porém que possui uma metodologia distinta do agrupamento hierárquico). Os resultados mostraram que o k-means teve um melhor desempenho que o agrupamento hierárquico não espacial. Já em uma comparação do k-means com o agrupamento hierárquico espacial, entendeu-se que a incorporação de contiguidade trouxe contribuições bastante relevantes.

Como apresentado, existem diversos estudos na literatura que abordam a distribuição espacial de diferentes doenças, comprovando que existe relação na ocorrência de doenças com o espaço geográfico e essa é uma área, ao mesmo tempo, com muitos conhecimentos e também com muitas descobertas a serem feitas, pois muitos desses estudos possuem foco em um município ou estado específico. Porém, não foram encontrados estudos que buscassem associar todos os tipos de doenças a fim de identificar quais possuem comportamentos semelhantes, e é essa a grande motivação deste trabalho. Além disso, nos artigos mencionados há uma grande diversidade de metodologias adotadas, ou seja, não é trivial a definição de uma técnica a ser utilizada em análises espaciais. Visto isso, a metodologia adotada nesta pesquisa testou algumas das técnicas apresentadas nos estudos acima (os resultados negativos estão detalhadas no Capítulo 5) e adotou uma metodologia própria, que está descrita no próximo capítulo, junto com os dados utilizados e as análises preliminares que apoiaram sua construção.

Capítulo 4

Metodologia

Neste capítulo, descrevemos o processo adotado para atingir os dois grandes objetivos deste trabalho. A Seção 4.1 resume todos os processos que constituem a metodologia, a Seção 4.2 referencia de onde e como os dados foram extraídos, a Seção 4.3 apresenta algumas análises preliminares essenciais para o desenvolvimento do trabalho, a Seção 4.4 demonstra todo o processo de preparação dos dados antes da aplicação do algoritmo de agrupamento hierárquico e, por fim, a Seção 4.5 define como foi aplicado o algoritmo de aprendizado de máquina e como foram construídas as análises dos *clusters* criados.

Todas as extrações e análises foram realizadas através do Jupyter Notebook[1], na linguagem Python, e as bibliotecas estão descritas ao longo do texto. Os códigos construídos, dados utilizados e imagens geradas estão disponíveis em repositórios do GitLab[2].

4.1 Etapas da metodologia

A metodologia deste trabalho contempla quatro grandes etapas, ilustradas na Figura 4.1.

1. Extração e consolidação dos dados;
2. Preparação dos dados;
3. Aplicação do algoritmo de aprendizado de máquina;
4. Análise dos resultados obtidos.

A extração dos dados é feita separadamente para três conjuntos de dados: óbitos, estimativas populacionais e dados geoespaciais dos distritos administrativos da cidade de São Paulo. Cada conjunto de dados extraído tem as suas especificidades e requerem alguns tratamentos, como remoção de linhas e colunas irrelevantes, substituições de valores ausentes por zero e alteração nos nomes dos distritos administrativos para todos os dados

[1] https://jupyter.org/

[2] https://gitlab.com/interscity/health/geospatial-analysis-of-deaths/-/tree/main/

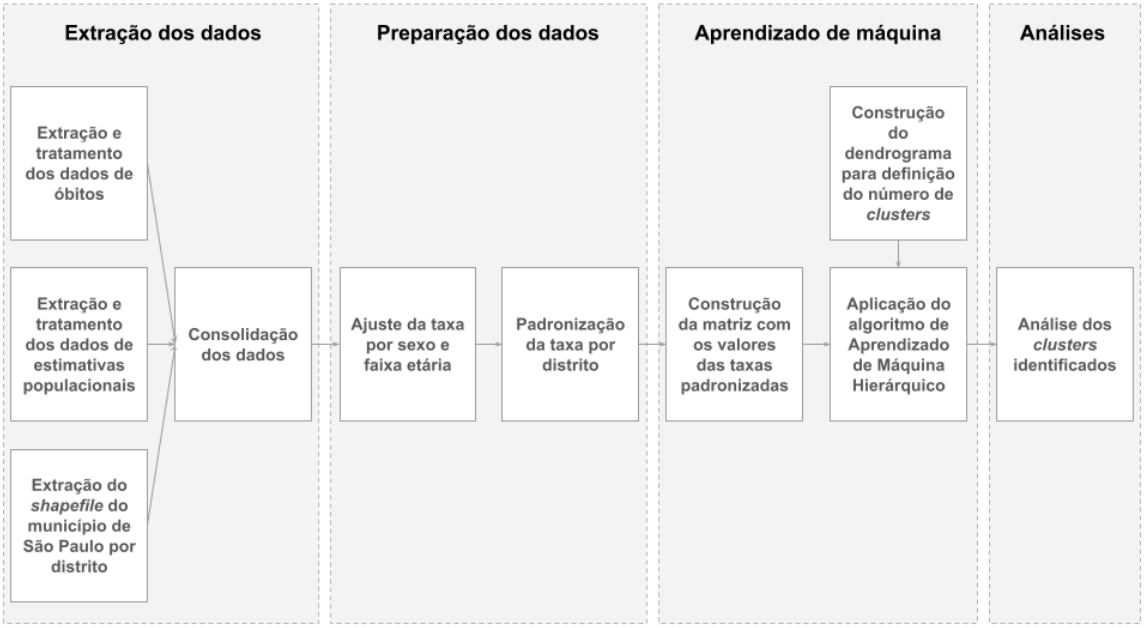


Figura 4.1: *Etapas da metodologia.*

serem compatíveis. Ao fim da extração e tratamento desses dados, eles são consolidados em um único arquivo e seguem para a próxima etapa.

Durante a preparação dos dados, são calculadas as taxas de mortalidade ajustadas por sexo e faixa etária, e então, são calculadas as taxas das doenças padronizadas por distrito administrativo, conforme demonstrado na Seção 2.1.

Na etapa de aprendizado de máquina é construída a matriz com os valores das taxas finais que serão utilizadas no algoritmo de agrupamento hierárquico. O dendrograma (conforme descrito na Seção 2.2) também é construído para auxiliar na definição do número ideal de *clusters* e então, o algoritmo é aplicado para definir os *clusters* finais.

Com os *clusters* definidos, são realizadas algumas análises espaciais e descritivas com o objetivo de interpretar os resultados obtidos e concluir quais são os grupos de doenças que possuem distribuições geográficas similares e quais distritos administrativos possuem comportamentos similares em relação às doenças causadoras de morte. Para ambas as questões de pesquisa, a metodologia é a mesma, mudando apenas o formato da matriz com os valores das taxas finais. Nessa última etapa, também são calculados os coeficientes de silhueta dos modelos, conforme descrito na Seção 2.3, para verificar se os grupos identificados estão adequados.

4.2 Dados utilizados

Foram utilizados os dados de mortalidade geral da cidade de São Paulo por distrito administrativo de residência e por capítulo da CID 10 associado à doença causadora do óbito. Portanto, não estão sendo considerados aqueles óbitos ocorridos em São Paulo, porém de indivíduos que não residem na cidade. Vale destacar que também foram testadas neste

trabalho metodologias com a própria CID 10 ao invés dos capítulos, porém devido à grande quantidade de doenças e, consequentemente, ao baixo número de mortes em cada uma delas (e muitos *outliers*) optou-se por trabalhar com os capítulos para obter-se maior consistência nos resultados. Um detalhamento de como se distribuem os distritos administrativos pelo município pode ser encontrado no Apêndice A e a relação dos capítulos com os códigos da CID 10 estão listados na Tabela 4.1. Foram extraídos os dados de 2014 a 2018, uma vez que a partir de 2019 os dados ainda não haviam sido disponibilizados nas granularidades necessárias no início deste trabalho.

Capítulo	Descrição	Códigos da CID 10
I	Algumas doenças infecciosas e parasitárias	A00-B99
II	Neoplasmas [tumores]	C00-D48
III	Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários	D50-D89
IV	Doenças endócrinas, nutricionais e metabólicas	E00-E90
V	Transtornos mentais e comportamentais	F00-F99
VI	Doenças do sistema nervoso	G00-G99
VII	Doenças do olho e anexos	H00-H59
VIII	Doenças do ouvido e da apófise mastóide	H60-H95
IX	Doenças do aparelho circulatório	I00-I99
X	Doenças do aparelho respiratório	J00-J99
XI	Doenças do aparelho digestivo	K00-K93
XII	Doenças da pele e do tecido subcutâneo	L00-L99
XIII	Doenças do sistema osteomuscular e do tecido conjuntivo	M00-M99
XIV	Doenças do aparelho geniturinário	N00-N99
XV	Gravidez, parto e puerpério	O00-O99
XVI	Algumas afecções originadas no período perinatal	P00-P96
XVII	Malformações congênitas, deformidades e anomalias cromossômicas	Q00-Q99
XVIII	Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte	R00-R99
XIX	Lesões, envenenamentos e algumas outras conseqüências de causas externas	S00-T98
XX	Causas externas de morbidade e de mortalidade	V01-Y98
XXI	Fatores que influenciam o estado de saúde e o contato com os serviços de saúde	Z00-Z99
XXII	Códigos para propósitos especiais	U00-U99
**	CID 10ª Revisão não disponível ou não preenchido ou inválido	Em branco ou inválido

Tabela 4.1: *Detalhamento dos capítulos da CID 10. Fonte: CID10, 2013.*

Os dados de óbitos[3] foram obtidos através do TABNET, que é uma plataforma online

[3] http://tabnet.saude.prefeitura.sp.gov.br/cgi/deftohtm3.exe?secretarias/saude/TABNET/SIM/obito.def

de acesso público que permite ao usuário realizar consultas nos dados, para subsidiar análises e tomadas de decisão na área da saúde (Saúde, 2023). O aplicativo foi desenvolvido pelo DATASUS (Departamento de Informática do Sistema Único de Saúde) para extrair informações das bases de dados do Sistema Único de Saúde (SUS) e pode ser facilmente parametrizável, como apresentado na Figura 4.2. Nele, estão contemplados os dados de diversos sistemas do SUS, inclusive do Sistema de Informações sobre Mortalidade (SIM), que traz informações sobre todas as mortes ocorridas no Brasil detalhadas por diversas variáveis, como ano, doença, localidade, entre outras. Na coleta de dados, é possível definir a informação que irá na linha da tabela, na coluna, o conteúdo e o período. Feita a seleção desses campos, a tabela pode ser exportada para um arquivo no formato *csv* (valores separados por vírgulas).

Figura 4.2: *Aplicativo TABNET para extração de dados do SUS.*

Também é possível, através do aplicativo, adicionar filtros às extrações, como por exemplo, sexo e faixa etária. Para as análises preliminares, os dados foram extraídos considerando-se esses dois filtros. Portanto, foram extraídas dez tabelas para cada ano, como detalhado na Tabela 4.2, totalizando cinquenta arquivos em formato *csv*.

Sexo	Faixa Etária
Feminino	Até 19 anos
	20 a 39 anos
	40 a 59 anos
	60 a 74 anos
	75 anos ou mais
Masculino	Até 19 anos
	20 a 39 anos
	40 a 59 anos
	60 a 74 anos
	75 anos ou mais

Tabela 4.2: *Bases de dados extraídas pelo TABNET para cada ano.*

Além dos dados de mortalidade, também foram extraídas pelo TABNET as estimativas populacionais[4] por distrito administrativo para os mesmos anos (2014 a 2018), elaboradas pela Fundação SEADE (Sistema Estadual de Análise de Dados). Da mesma forma, esses dados foram extraídos segmentados por sexo e faixa etária, para que se pudesse calcular as taxas de mortalidade por 100 mil habitantes de cada uma dessas categorias. Calculadas as taxas para cada doença, distrito administrativo, sexo, faixa etária e ano, foram obtidas as médias dos cinco anos para cada uma dessas segmentações a fim de evitar conclusões enviesadas.

Por fim, o *shapefile*[5] dos distritos administrativos de São Paulo foi obtido através da Secretaria Municipal de Desenvolvimento Urbano (SMDU), com sua última atualização no ano de 2015. *Shapefile* é uma coleção de arquivos contendo informações geoespaciais de uma região (Dados Abertos, 2015) e foi utilizado neste trabalho para gerar os mapas da cidade de São Paulo.

Para a extração dos dados, foi construído um *webcrawler* utilizando a biblioteca *selenium*[6] do Python que, dadas as definições anteriores, automatizou a extração de todos os arquivos. Para o carregamento e tratamento dos dados de mortalidade e população foi utilizada a biblioteca *pandas*[7]. Já para a leitura do *shapefile* dos distritos administrativos e para a criação dos mapas, foi utilizada a biblioteca *geopandas*[8], específica para análises geoespaciais.

4.3 Análises preliminares

Como apresentado no Capítulo 3, diversos estudos mostraram diferentes intensidades na ocorrência das doenças para diferentes grupos de faixa etária e sexo. Portanto, uma análise preliminar foi construída para entender se há necessidade desse estudo ser segmentado para diferentes populações de acordo com essas duas variáveis. Os gráficos apresentados nesta seção foram construídos com o auxílio da biblioteca *plotly*[9] do Python.

Alguns distritos administrativos da cidade de São Paulo possuem uma fração da população idosa (60 anos ou mais) muito maior que outros, sobretudo a região central, como pode ser observado na Figura 4.3a. Portanto, primeiramente buscou-se entender a influência desse perfil na taxa de mortalidade, que está demonstrada na Figura 4.3b. Nessa análise, foi construído um gráfico de dispersão da taxa média de mortalidade por 100 mil habitantes *versus* a fração da população idosa por distrito administrativo. Esse gráfico pode ser visualizado pela Figura 4.4. Nele, cada ponto representa um distrito administrativo e os tamanhos representam a população total daquele distrito. Pelo gráfico, pode-se observar que a variação tanto da fração da população idosa quanto da taxa de mortalidade é bastante elevada, variando mais de 3 vezes. Além disso, é possível observar

[4] http://tabnet.saude.prefeitura.sp.gov.br/cgi/deftohtm3.exe?secretarias/saude/TABNET/POP/pop.def

[5] http://dados.prefeitura.sp.gov.br/pt_PT/dataset/distritos/resource/9e75c2f7-5729-4398-8a83-b4640f072b5d

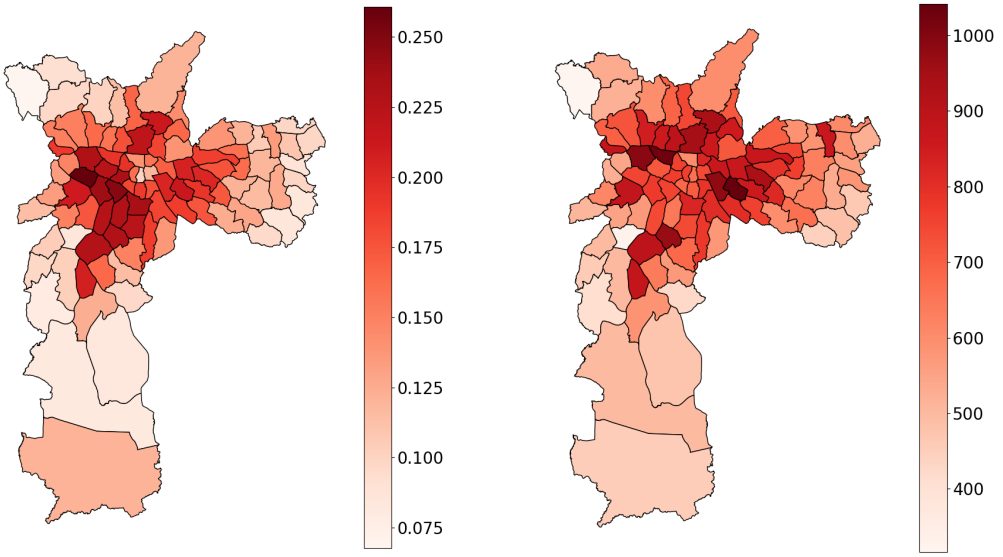
[6] https://pypi.org/project/selenium/

[7] https://pandas.pydata.org/

[8] https://geopandas.org/en/stable/

[9] https://plotly.com/python/

que quanto maior a fração da população idosa em um distrito administrativo, maior é a taxa de mortalidade geral e essa relação pode ser ajustada por uma curva logarítmica (indicada no gráfico).



(a) *Fração da população idosa.* **(b)** *Taxa de mortalidade por 100 mil habitantes.*

Figura 4.3: *Distribuições da fração da população idosa e taxa de mortalidade por distrito administrativo.*

Para uma melhor visualização, foi calculada a diferença de cada ponto no gráfico anterior para a curva logarítmica projetada e essa diferença foi representada em um mapa coroplético (Figura 4.5). Nesse mapa, quanto maior a intensidade do verde, maior é a diferença e essa diferença é negativa, isso significa que a taxa de mortalidade observada é menor que a taxa projetada pela curva. Quanto maior a intensidade do vermelho, maior é a diferença e essa diferença é positiva, isso significa que a taxa de mortalidade observada é maior que a taxa projetada pela curva. Já os distritos administrativos em amarelo possuem uma taxa observada muito próxima à taxa projetada.

Através dessa imagem podemos ver que existem alguns distritos administrativos com uma fração da população idosa baixa, porém com altas taxas de mortalidade (como Barra Funda, São Miguel Paulista e Água Rasa, que são os três distritos com maiores intensidades de vermelho no mapa) e também alguns com uma fração da população idosa alta, mas com baixas taxas de mortalidade (como Moema, Morumbi e Alto de Pinheiros, que são os três distritos com maiores intensidades de verde no mapa). A Figura A.1 pode auxiliar na localização desses distritos com maior facilidade. Para entender se esses *outliers* podem ser explicados pela situação socioeconômica desses distritos, foi calculada a correlação entre a diferença apresentada no mapa e o Índice de Desenvolvimento Humano (IDH) dos distritos, segundo o censo de 2000, e obteve-se uma correlação negativa e fraca (-0,17). O IDH de cada distrito administrativo pode ser observado na Figura 4.6 e vale destacar que esse índice é maior na região central da cidade. Portanto, distritos com piores qualidades de vida podem apresentar uma diferença maior, porém vale ressaltar que é uma correlação bastante fraca. Visto isso, essas conclusões podem ser bastante relevantes para gestores de saúde, uma vez que, apesar de existir uma relação bastante forte entre a taxa de mortalidade e a

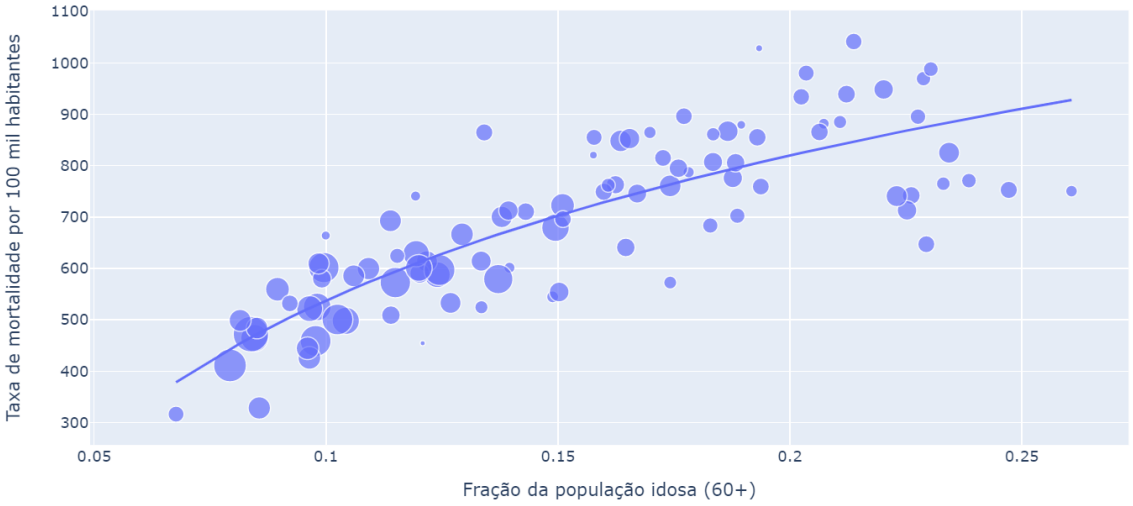


Figura 4.4: *Dispersão da taxa de mortalidade por 100 mil habitantes versus população idosa por distrito administrativo.*

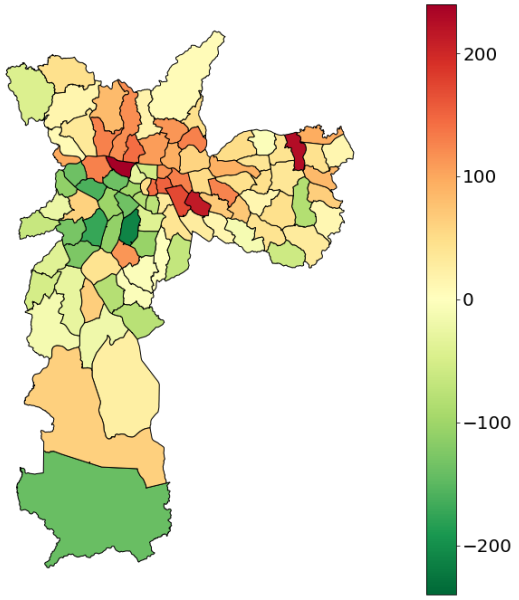


Figura 4.5: *Diferença da taxa de mortalidade para a curva logarítmica. Distritos em verde possuem uma diferença negativa entre a taxa de mortalidade e a taxa projetada pela curva logarítmica*[a]*. Distritos em vermelho possuem uma diferença positiva e, aqueles em amarelo, possuem uma diferença próxima de zero.*

[a] O distrito de Marsilac, localizado no extremo sul da cidade, pode ser considerado um *outlier* devido à baixa população e ao baixo número de óbitos. Portanto, temos uma menor segurança estatística neste resultado.

fração de população idosa nos distritos, existem alguns destaques onde há poucos idosos, porém uma alta taxa de mortalidade. Isso pode ser um indício de um problema localizado nessas regiões, como a falta de Unidades Básicas de Saúde (UBS), por exemplo. Vale uma investigação mais profunda de qual pode ser o problema que afeta essas regiões, caso ainda não seja um problema conhecido pelos gestores. Por outro lado, também existem destaques de distritos que possuem uma população idosa bastante relevante, mas com baixas taxas de mortalidade. Esses podem servir como estudo de caso, para que se possa entender o que contribuiu para esse bom resultado, e assim, prover políticas de melhorias para os demais.

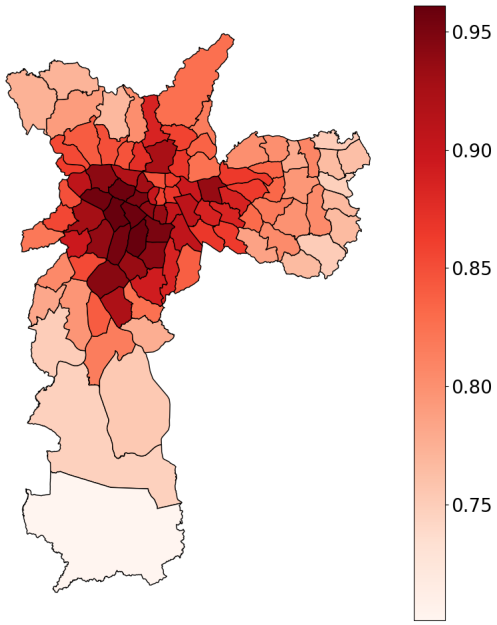
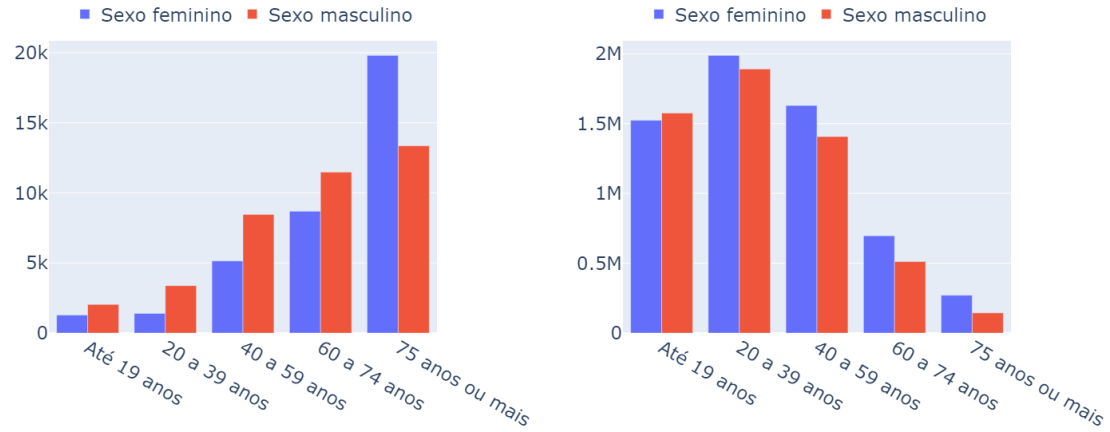


Figura 4.6: *Índice de Desenvolvimento Humano por distrito administrativo.*

Foram construídos também com os dados extraídos, gráficos das distribuições de óbitos e população médios por ano (de 2014 a 2018) nos grupos de sexo e faixa etária, conforme ilustrado nas Figuras 4.7a e 4.7b, indicando que a distribuição desses dois indicadores seguem diferentes padrões. Quanto maior a faixa etária da população, maior o número de óbitos. Esse número é maior para o sexo masculino em todas as faixas etárias, exceto para 75 anos ou mais, indicando que a maior parte da população desse sexo vai a óbito com idade inferior à população feminina. Já em relação à população, vemos um comportamento inverso. Ela é crescente da faixa de até 19 anos para a faixa de 20 a 39 anos e decrescente a partir de então. A população feminina é superior em todas as faixas etárias, com exceção da faixa de até 19 anos onde a população masculina é um pouco maior.

Após essas análises preliminares, concluiu-se que a identificação dos grupos de doenças deve levar em consideração as especificidades de sexo e faixa etária. Sendo assim, utilizou-se o ajuste por sexo e faixa etária definido anteriormente para construir uma taxa de mortalidade única que leve em consideração essa necessidade. Porém, como o ajuste define no seu cálculo pesos maiores para populações maiores, entendeu-se que a utilização das faixas etárias abaixo de 40 anos poderiam enviesar os dados. Isto porque as duas primeiras faixas etárias analisadas possuem populações grandes e um número baixo de óbitos. Sendo



(a) *Distribuição dos óbitos (médias por ano).* **(b)** *Distribuição da população (médias por ano).*

Figura 4.7: *Distribuições por sexo e faixa etária da população total e dos óbitos do município de São Paulo.*

assim, a taxa de mortalidade ajustada teria grande influência dessas faixas etárias e não seria representativa em relação aos óbitos. Visto isso, este trabalho optou por utilizar apenas as faixas etárias a partir de 40 anos.

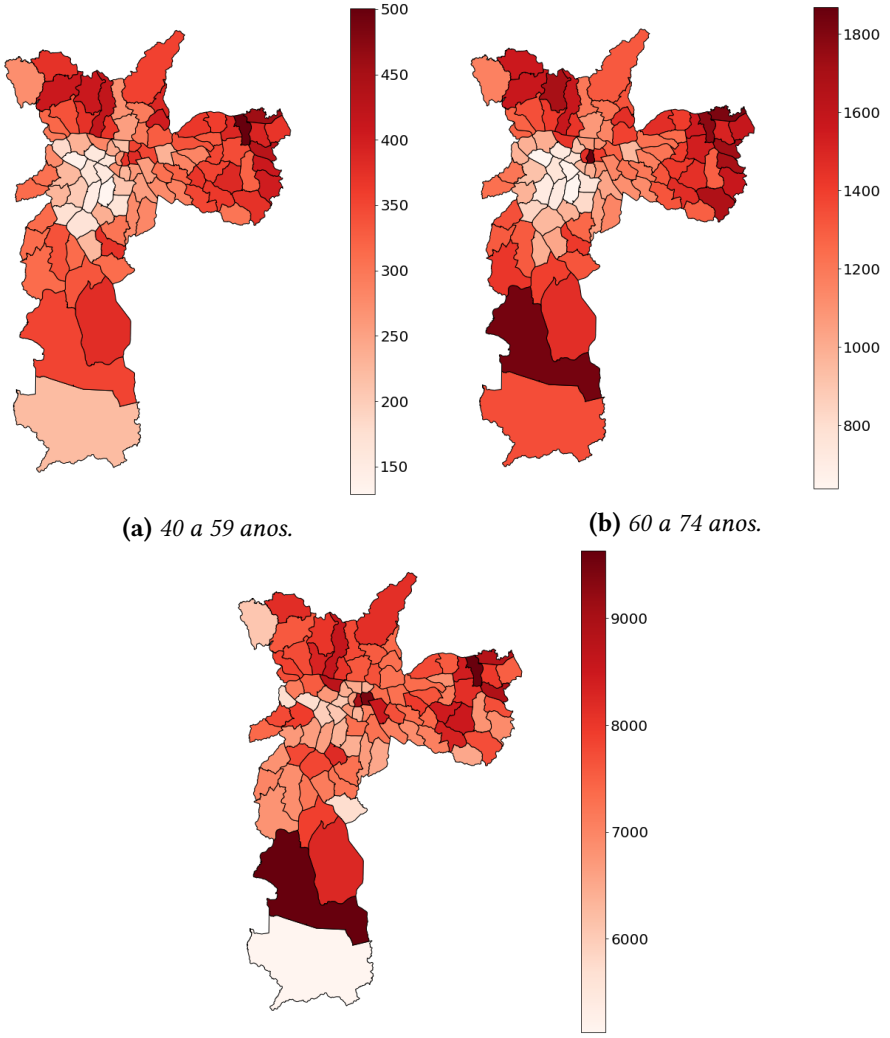
Uma outra modificação realizada nas categorias de faixa etária foi o agrupamento das faixas de 40 a 59 anos e 60 a 74 anos, pois ambas possuem distribuições geográficas semelhantes, como pode ser visto nas Figuras 4.8 e 4.9. Para confirmação dessa percepção também foram calculadas as distâncias euclidianas entre os vetores das taxas de mortalidade que construíram cada um desses mapas e os resultados obtidos estão demonstrados na Tabela 4.3. De fato, podemos comprovar que as distâncias entre os vetores das faixas etárias de 40 a 59 anos e 60 a 74 anos são as menores, para ambos os sexos.

Sexo	Faixa Etária (Vetor 1)	Faixa Etária (Vetor 2)	Distância
Feminino	40 a 59 anos	60 a 74 anos	9.466
	40 a 59 anos	75 anos ou mais	70.060
	60 a 74 anos	75 anos ou mais	60.792
Masculino	40 a 59 anos	60 a 74 anos	16.658
	40 a 59 anos	75 anos ou mais	86.361
	60 a 74 anos	75 anos ou mais	69.962

Tabela 4.3: *Distâncias calculadas entre os vetores de taxa de mortalidade de cada combinação de faixa etária para o sexo feminino e masculino.*

4.4 Preparação dos dados

A etapa de preparação dos dados foi realizada com o auxílio da biblioteca *pandas* do Python e seguiu as metodologias definidas no Capítulo 2. Primeiro, foi realizado o ajuste

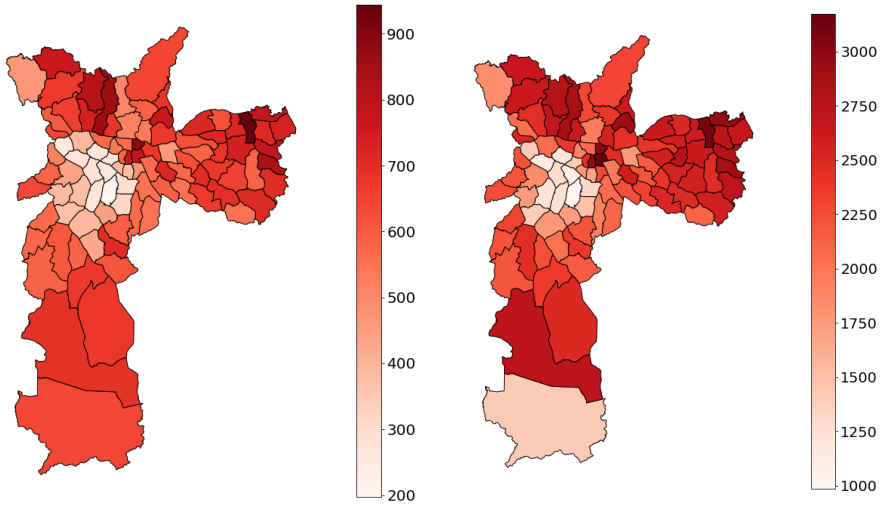


(a) *40 a 59 anos.*

(b) *60 a 74 anos.*

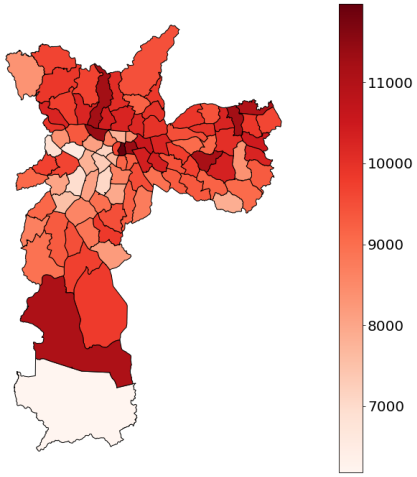
(c) *75 anos ou mais.*

Figura 4.8: *Distribuição das taxas de mortalidade da população do sexo feminino para cada faixa etária.*



(a) *40 a 59 anos.*

(b) *60 a 74 anos.*



(c) *75 anos ou mais.*

Figura 4.9: *Distribuição das taxas de mortalidade da população do sexo masculino para cada faixa etária.*

das taxas de mortalidade por 100 mil habitantes por sexo e faixa etária, utilizando-se como parâmetro as populações de duas faixas, de 40 a 74 anos e 75 anos ou mais, conforme concluído na seção anterior.

É importante ressaltar que a metodologia do ajuste adota uma única proporção populacional, que é a da cidade de São Paulo, e a aplica para todos os distritos administrativos. Ou seja, as taxas ajustadas consideram que todos os distritos possuem as mesmas proporções de população de cada categoria de sexo e faixa etária e, consequentemente, elas podem ser comparadas sem problemas de conclusões enviesadas.

Por fim, essa taxa ajustada foi padronizada por distrito administrativo. Essa padronização é realizada pois durante a comparação das doenças, é importante que os dados estejam na mesma escala dentro de um mesmo distrito administrativo. Dessa forma, comparações de várias doenças dentro de um distrito também podem ser realizadas sem problemas de conclusões enviesadas. Vale mencionar que também foram realizados testes com taxas

padronizadas por doença ao invés de distritos administrativos. Neste caso, os dados foram padronizados para estarem na mesma escala para cada doença. Porém, os testes realizados com as taxas padronizadas por distrito administrativos apresentaram melhores resultados após a aplicação do algorítimo de agrupamento e portanto, optou-se por utilizar essa padronização.

4.5 Análises

Com os dados prontos, a próxima etapa consiste na aplicação do algoritmo de agrupamento hierárquico. A matriz de dados utilizada pelo algoritmo varia de acordo com a questão de pesquisa que está sendo trabalhada. Para a primeira questão ("Quais grupos de doenças causadoras de morte na cidade de São Paulo possuem distribuições geográficas similares?"), a matriz é construída com cada linha representando um capítulo da CID 10, cada coluna representando um distrito administrativo e com o conteúdo sendo as taxas de mortalidade padronizadas finais, como demonstra a Figura 4.10a. A matriz é construída dessa forma pois queremos agrupar doenças de acordo com o vetor de mortalidade nos distritos para cada doença. Já para a segunda questão ("Quais distritos administrativos da cidade de São Paulo possuem comportamentos similares em relação às doenças causadoras de morte?"), a matriz é o transposto da matriz anterior. Ou seja, cada linha representa um distrito administrativo, cada coluna representa um capítulo da CID 10 e o conteúdo continua sendo as taxas de mortalidade padronizadas finais, como demonstra a Figura 4.10b. Neste caso, queremos agrupar os distritos administrativos de acordo com o vetor de mortalidade das doenças em cada distrito.

Capítulo CID 10	Água Rasa	Alto de Pinheiros	...	Vila Sônia
I. Algumas doenças infecciosas e parasitárias				
II. Neoplasias (tumores)		Taxas de mortalidade		
...				
XX. Causas externas de morbidade e mortalidade				

(a) *Primeira questão de pesquisa.*

Distrito Administrativo	I. Algumas doenças infecciosas e parasitárias	II. Neoplasias (tumores)	...	XX. Causas externas de morbidade e mortalidade
Água Rasa				
Alto de Pinheiros		Taxas de mortalidade		
...				
Vila Sônia				

(b) *Segunda questão de pesquisa.*

Figura 4.10: *Matrizes construídas para responder às duas questões de pesquisa.*

Com o auxílio do dendrograma é definido o número ideal de *clusters* e, então, o algoritmo de agrupamento hierárquico é aplicado. No Capítulo 5 detalharemos os testes realizados com os demais algoritmos de aprendizado não supervisionado descritos no Capítulo 2, justificando a escolha pelo agrupamento hierárquico. Ao final desse processo, temos definido quais capítulos da CID 10 (ou quais distritos administrativos) pertencem a um mesmo *cluster*.

Para a primeira questão de pesquisa, como estamos agrupando doenças, as análises

finais consistem no cálculo das taxas médias de mortalidade de todas as doenças pertencentes a cada *cluster* final para cada distrito administrativo, para que se possa criar um único mapa da distribuição da taxa de mortalidade por *cluster* e então compará-los. Já para a segunda questão de pesquisa, como estamos agrupando distritos administrativos, conseguimos criar um único mapa para distinguir quais são os distritos que pertencem a cada um dos *clusters* identificados e, então, comparar as taxas médias de mortalidade de cada doença em cada um dos grupos. Os gráficos e mapas criados nesta etapa serão apresentados no Capítulo 5. Por fim, também são analisados os coeficientes de silhueta de cada modelo, para que se tenha uma ideia do quão ideal estão os *clusters* identificados.

O próximo capítulo apresenta algumas estatísticas descritivas dos dados, alguns resultados negativos obtidos antes de se definir a metodologia apresentada neste capítulo, os *clusters* finais obtidos sob as duas diferentes perspectivas definidas no Capítulo 1 e as análises espaciais e descritivas que auxiliam na interpretação dos resultados obtidos.

Capítulo 5

Resultados

Este capítulo tem como objetivo apresentar os resultados finais obtidos após a aplicação da metodologia definida no Capítulo 4 e também alguns resultados negativos obtidos antes que fosse estabelecida a metodologia final. A Seção 5.1 traz alguns números gerais do conjunto de dados e também algumas análises descritivas espaciais. A Seção 5.2 mostra análises realizadas que não obtiveram resultados satisfatórios, mas que incluímos para servir de aprendizado para pesquisas futuras. A Seção 5.3 apresenta os resultados que auxiliam na resposta da primeira questão de pesquisa: "Quais grupos de doenças causadoras de morte na cidade de São Paulo possuem distribuições geográficas similares?". De forma análoga, a Seção 5.4 mostra os resultados que contribuem na resposta da segunda questão de pesquisa: "Quais distritos administrativos da cidade de São Paulo possuem comportamentos similares em relação às doenças causadoras de morte?". Ao final das seções 5.3 e 5.4 existem subseções para discussão dos resultados exibidos.

5.1 Estatística descritiva

São Paulo é um município com aproximadamente 1.521 km² de área e com uma população de 11.960.216 habitantes, segundo estimativa da fundação SEADE de 2022, sendo o município mais populoso do país (Geografia e Estatística, 2017). Nos anos de 2014 a 2018 foram registrados uma média de 75.293 óbitos por ano e a Figura 5.1 mostra que esse número não sofreu grandes alterações no período mencionado.

A cidade divide-se em 96 distritos administrativos e as suas áreas, estimativas populacionais médias e médias de óbitos por ano entre 2014 e 2018 estão disponíveis nas Tabelas A.1 a A.3 do Anexo A. Os mapas das Figuras 5.2 e 5.3 possibilitam uma melhor visualização dos valores apresentados. As tabelas e mapas mostram os valores tanto considerando a população total, quanto considerando a população acima de 40 anos, que é o público alvo desse estudo. Alguns destaques interessantes são:

- O distrito Marsilac possui a maior área territorial e também a menor população e a menor quantidade de óbitos. Esse é um destaque importante pois mostraremos mais a seguir que esses números podem produzir alguns *outliers* nos resultados.

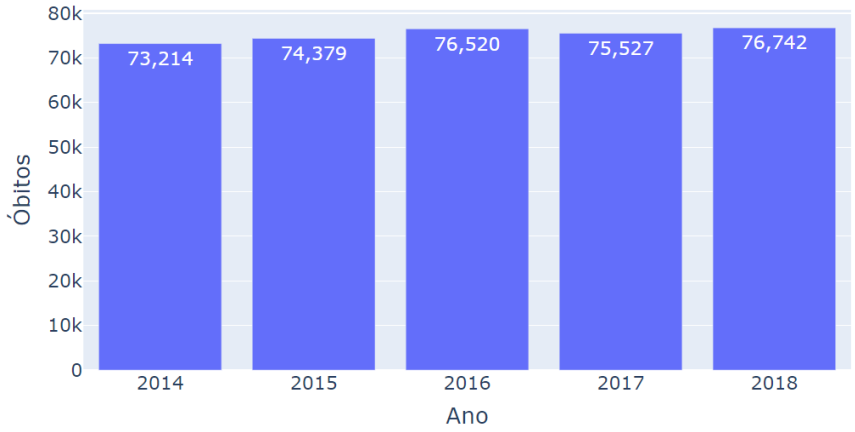
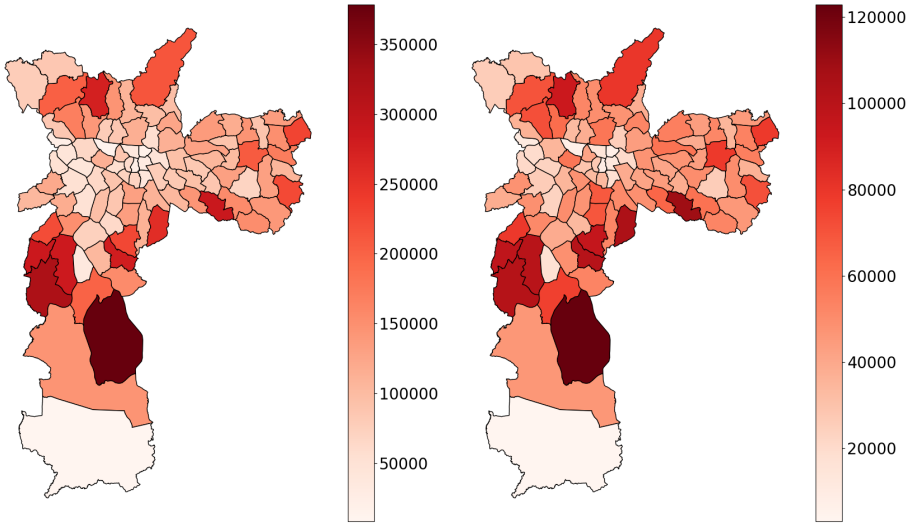


Figura 5.1: *Óbitos relativos aos residentes da cidade de São Paulo entre 2014 e 2018.*

- O distrito com maior população e maior quantidade de óbitos é o Grajaú, localizado na zona sul da cidade.
- Sapopemba é o segundo distrito com a maior quantidade de óbitos total, porém é o distrito com a maior quantidade de óbitos da população com 40 anos ou mais.

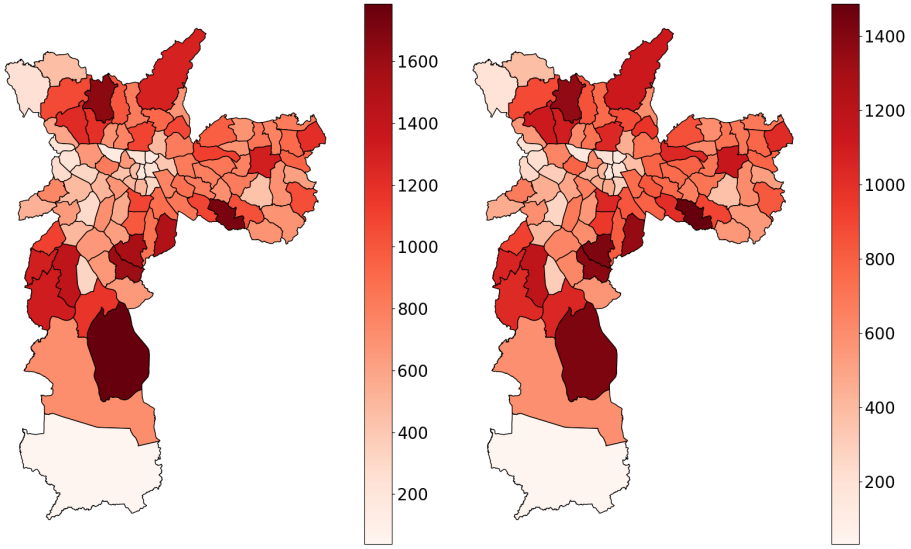
Nas figuras podemos ver com mais clareza que a população e os óbitos, tanto total quanto acima de 40 anos, estão mais concentrados na região periférica da cidade. Porém, quando analisamos as taxas de mortalidade por 100 mil habitantes (Figura 5.4), vemos uma maior concentração na região central. Essas conclusões ressaltam a importância de trabalharmos com as taxas de mortalidade ao invés dos valores brutos de óbitos, pois é natural que distritos mais populosos terão maiores números de óbitos, mas não necessariamente possuirão as maiores taxas.



(a) *População total.* **(b)** *População acima de 40 anos.*

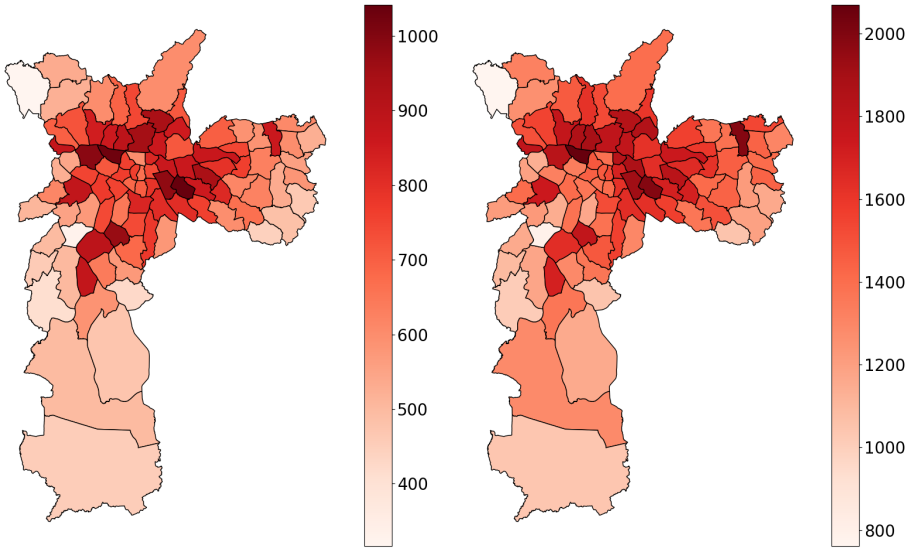
Figura 5.2: *Médias das estimativas populacionais.*

Também foram construídas análises descritivas para os dados de acordo com os capítulos da CID 10. Nos dados, não apareceram óbitos para os capítulos XIX (lesões, en-



(a) *População total.* **(b)** *População acima de 40 anos.*

Figura 5.3: *Médias dos óbitos.*



(a) *População total.* **(b)** *População acima de 40 anos.*

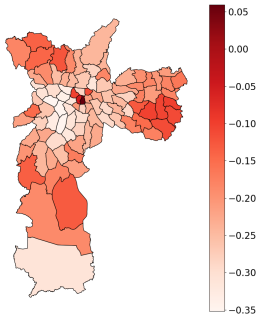
Figura 5.4: *Taxa de mortalidade por 100 mil habitantes.*

venenamentos e algumas outras conseqüências de causas externas), XXI (fatores que influenciam o estado de saúde e o contato com os serviços de saúde) e XXII (códigos para propósitos especiais) nos anos observados. As quantidades médias de óbitos por ano por capítulo da CID 10, tanto para a população total quanto para a população acima de 40 anos, podem ser observadas na Tabela 5.1. Para ambas as populações, mais da metade dos óbitos estão concentrados nos capítulos IX (doenças do aparelho circulatório) (32% para a população total e 35% para a população acima de 40 anos) e II (Neoplasmas [tumores]) (20% e 22%, respectivamente). Pouquíssimos óbitos foram registrados por ano nos capítulos VII (doenças do olho e anexos) e VIII (doenças do ouvido e da apófise mastóide). Além desses dois, especificamente para a população acima de 40 anos, como esperado, também foram registrados poucos óbitos para os capítulos XV (gravidez, parto e puerpério), XVI (algumas afecções originadas no período perinatal) e XVII (malformações congênitas, deformidades e anomalias cromossômicas). Visto isso, com o intuito de remover *outliers*, esses cinco capítulos foram excluídos dos dados de análise. Os mapas com as taxas de mortalidade padronizadas por capítulo da CID 10 podem ser observados nas Figuras 5.5 e 5.6.

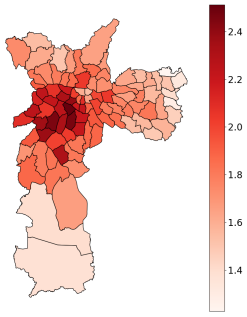
Capitulo CID 10	Óbitos (total)	Óbitos (40+)
I. Algumas doenças infecciosas e parasitárias	2.752	2.246
II. Neoplasmas [tumores]	15.322	14.597
III. Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários	231	175
IV. Doenças endócrinas, nutricionais e metabólicas	3.135	3.008
V. Transtornos mentais e comportamentais	1.149	1.102
VI. Doenças do sistema nervoso	2.869	2.598
VII. Doenças do olho e anexos*	1	1
VIII. Doenças do ouvido e da apófise mastóide*	12	10
IX. Doenças do aparelho circulatório	24.350	23.444
X. Doenças do aparelho respiratório	10.203	9.599
XI. Doenças do aparelho digestivo	4.157	3.898
XII. Doenças da pele e do tecido subcutâneo	404	391
XIII. Doenças do sistema osteomuscular e do tecido conjuntivo	375	332
XIV. Doenças do aparelho geniturinário	2.569	2.496
XV. Gravidez, parto e puerpério*	92	8
XVI. Algumas afecções originadas no período perinatal*	1.099	1
XVII. Malformações congênitas, deformidades e anomalias cromossômicas*	617	59
XVIII. Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte	792	479
XX. Causas externas de morbidade e de mortalidade	5.992	3.095

Tabela 5.1: *Média de óbitos por ano por capítulo da CID 10. *Capítulos excluídos da análise devido à baixa representatividade no número de óbitos.*

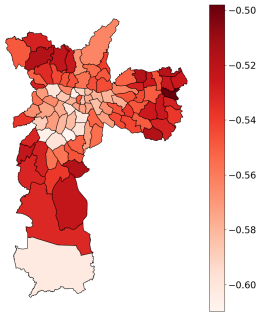
Em relação aos grupos de população que estão sendo trabalhados nesta pesquisa, segmentados por sexo e faixa etária, a Tabela 5.2 traz as médias das estimativas populacionais



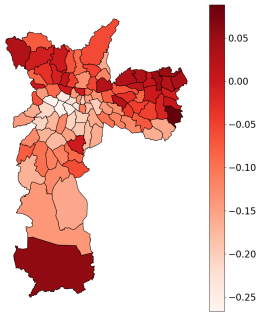
(a) *I. Algumas doenças infecciosas e parasitárias.*



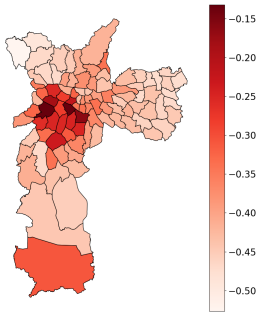
(b) *II. Neoplasmas [tumores].*



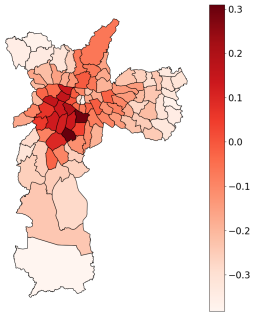
(c) *III. Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários.*



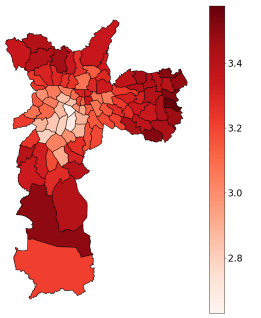
(d) *IV. Doenças endócrinas, nutricionais e metabólicas.*



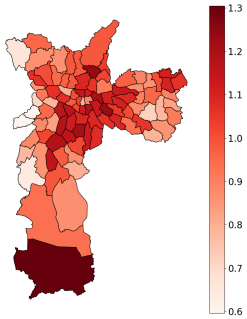
(e) *V. Transtornos mentais e comportamentais.*



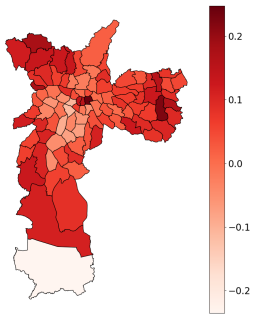
(f) *VI. Doenças do sistema nervoso.*



(g) *IX. Doenças do aparelho circulatório.*

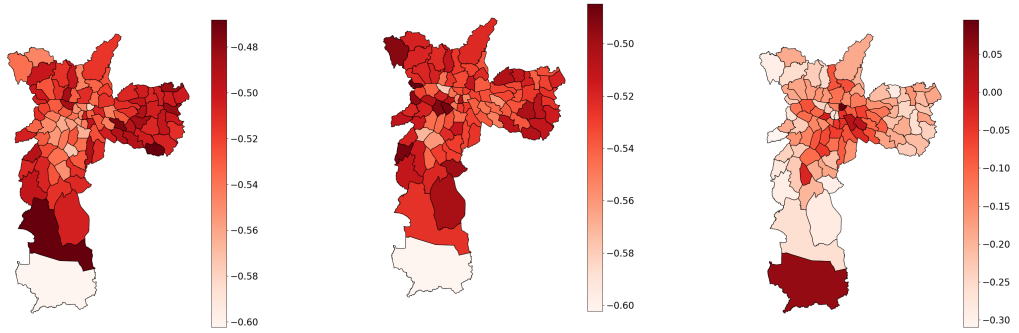


(h) *X. Doenças do aparelho respiratório.*



(i) *XI. Doenças do aparelho digestivo.*

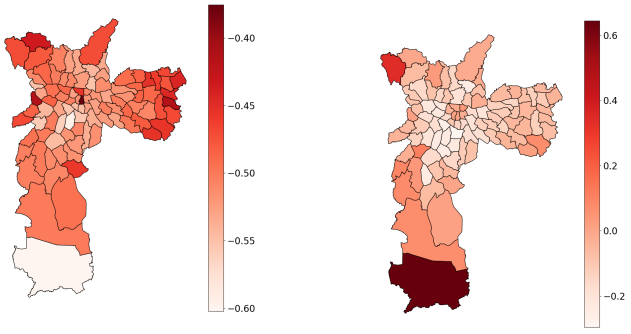
Figura 5.5: *Mapas das taxas de mortalidade por capítulo da CID 10 (1/2).*



(a) *XII. Doenças da pele e do tecido subcutâneo.*

(b) *XIII. Doenças do sistema osteomuscular e do tecido conjuntivo.*

(c) *XIV. Doenças do aparelho geniturinário.*



(d) *XVIII. Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte.*

(e) *XX. Causas externas de morbidade e de mortalidade.*

Figura 5.6: *Mapas das taxas de mortalidade por capítulo da CID 10 (2/2).*

e as médias de óbitos por ano entre 2014 e 2018, além das taxas de mortalidade por 100 mil habitantes. A tabela mostra que a maior parte dessa população é do sexo feminino com 40 a 74 anos e a menor população é do sexo masculino com 75 anos ou mais. Em relação aos números absolutos de óbitos, a maior concentração é na população masculina com 40 a 74 anos e a menor é na população masculina com 75 anos ou mais. Em relação às taxas de mortalidade, são mais representativas nas faixas de 75 anos ou mais, sendo maior para o sexo masculino.

Portanto, essa pesquisa abrangeu 96 distritos administrativos e 14 capítulos da CID 10, analisando óbitos da população acima de 40 anos para ambos os sexos feminino e masculino.

5.2 Resultados negativos

Antes de ser estabelecida a metodologia final apresentada, diversos testes foram realizados com o objetivo de responder às questões de pesquisa definidas. Apesar de não terem

Sexo	Faixa etária	Estimativa populacional	Óbitos	Taxa de mortalidade (por 100 mil habitantes)
Feminino	40 a 74 anos	2.326.096	13.910	598
Feminino	75 anos ou mais	270.176	19.868	7.354
Masculino	40 a 74 anos	1.918.612	20.348	1.061
Masculino	75 anos ou mais	144.003	13.412	9.314

Tabela 5.2: *Estimativa populacional média e média de óbitos por ano por sexo e faixa etária.*

produzido resultados satisfatórios, é importante destacarmos quais foram esses testes para servir de aprendizado para trabalhos futuros.

Como mencionado no Capítulo 4, inicialmente decidimos trabalhar com as doenças da CID 10 (agrupadas pelos três primeiros dígitos) ao invés dos capítulos. As primeiras análises já apontaram muitos *outliers*, uma vez que quando trabalhamos com aproximadamente mil categorias distintas, os óbitos ficam muito distribuídos e poucas doenças apresentaram uma quantidade significativa para que os resultados não fossem enviesados. Em conversas com alguns especialistas da área da saúde, foi sugerido trabalharmos com as doenças agrupadas pelos capítulos.

Muitos outros testes foram realizados utilizando o Índice Global de Moran e o Índice Local de Moran. Esses resultados estão apresentados na Seção 5.2.1. Além desses, um outro teste foi realizado de alteração do conteúdo das matrizes criadas. Dessa vez, optamos por utilizar tanto a taxa de mortalidade bruta quanto a taxa de mortalidade suavizada, para trazer uma abordagem espacial para o algoritmo de agrupamento. A suavização da taxa foi realizada através do estimador Bayesiano empírico local, apresentado também no trabalho de Costa-Nobre *et al.*, 2021, que recalcula as taxas de um distrito administrativo atribuindo um peso às taxas dos distritos vizinhos. Ambas produziram bons resultados e, como a taxa suavizada não trouxe incremento no coeficiente de silhueta, optamos por trabalhar com a taxa bruta. Os resultados com esse último teste são apresentados nas Seções 5.3 e 5.4.

5.2.1 Índice Global e Local de Moran

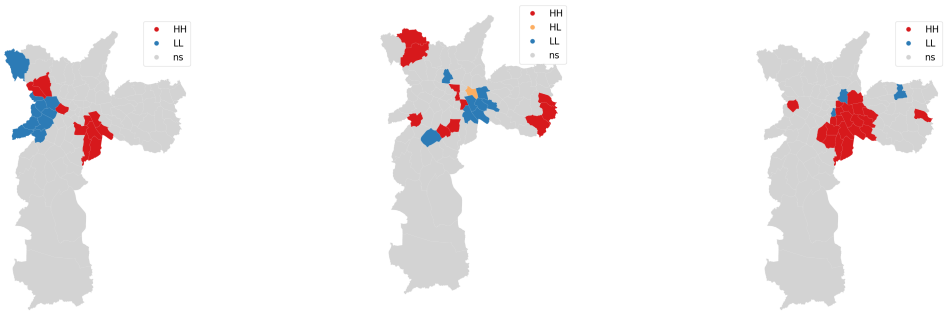
Durante a pesquisa de trabalhos relacionados, foram identificados muitos estudos que utilizaram o Índice de Moran para a obtenção de *clusters* espaciais, sendo uma técnica bastante disseminada na literatura. Visto isso, essa pesquisa dedicou bastante tempo na exploração dessa metodologia. Em um primeiro momento, foi utilizado o Índice Global de Moran para definir quais doenças possuem dependência espacial para, então, aplicar o Índice Local de Moran apenas para aquelas com dependência espacial. Também através de conversas com especialistas, entendemos que podem existir casos onde o Índice Global de Moran não aponta dependência espacial, porém é possível identificar *clusters* locais. Sendo assim, optou-se por trabalhar com o Índice Local de Moran para todas as doenças, deixando de lado a aplicação do índice global.

Assim como a metodologia final, os testes com o Índice Local de Moran consistiram na criação de uma matriz de dados para a aplicação do algoritmo de agrupamento. A matriz

foi definida da mesma forma como mencionado no Capítulo 4, ou seja, com as linhas contendo as doenças e as colunas contendo os distritos administrativos para a primeira questão de pesquisa, e com as linhas contendo os distritos administrativos e as colunas contendo as doenças para a segunda questão de pesquisa. A grande diferença se dá no conteúdo da matriz. Neste caso, foi testado o conteúdo das matrizes como sendo os LISA *Clusters* obtidos através dos índices locais calculados. Assim, as matrizes foram preenchidas com valores de 1 a 5, onde cada número representa uma das categorias mencionadas no Capítulo 2 (alto-alto, alto-baixo, baixo-alto, baixo-baixo e não significativo).

Para o LISA *Cluster* foram testados os algoritmos de agrupamento K-Means, K-Modes e também o agrupamento hierárquico, descritos no Capítulo 2. Como estamos trabalhando com valores que embora sejam numéricos, também podem ser analisados como categóricos, os três algoritmos foram testados. Tanto o K-Means quando o agrupamento hierárquico, produziram resultados muito similares, inclusive para o coeficiente de silhueta. Já o K-Modes produziu resultados menos satisfatórios. Como o algoritmo de agrupamento hierárquico já estava sendo aplicado para a construção do dendrograma, optou-se por seguir os testes com esse método.

Apesar de os resultados terem produzido um coeficiente de silhueta positivo, o valor ficou muito próximo de 0, indicando que os grupos criados tinham pouca distinção entre si. Em análises posteriores entendeu-se que grande parte dos LISA *Clusters* identificados estavam com valor igual a 5, ou seja, eram não significativos. Visto isso, o algoritmo criou um agrupamento enviesado pelos valores não significativos, ao invés de considerar os valores do LISA, que eram raros nas matrizes calculadas. A Figura 5.7 mostra um exemplo dos mapas dos LISA *Clusters* para três capítulos da CID 10 que foram agrupados em um mesmo *cluster* pelo algoritmo. Neles, vemos que as áreas de destaque como alto-alto (HH, em vermelho), alto-baixo (HL, em laranja), ou baixo-baixo (LL, em azul) são distintas entre os três mapas e o que eles têm em comum é a área destacada em cinza, ou seja, regiões com valores não significativos para o LISA. Para esse teste em específico, o coeficiente de silhueta foi de 0,20.



(a) *X. Doenças do aparelho respiratório.*

(b) *XIII. Doenças do sistema osteomuscular e do tecido conjuntivo.*

(c) *XIV. Doenças do aparelho geniturinário.*

Figura 5.7: *Mapas com os LISA Clusters de capítulos da CID 10 classificados em um mesmo cluster.*

Para mitigar esse problema, calculamos novamente os LISA *Clusters*, porém com a

taxa suavizada pelo estimador empírico de Bayes ao invés das taxas brutas, conforme foi realizado nos estudos de Costa-Nobre *et al.*, 2021 e Wang *et al.*, 2012. Esse estimador calcula as taxas de mortalidade atribuindo um peso ao distrito em análise e também um peso menor aos vizinhos de acordo com a matriz de proximidade espacial criada (Santos *et al.*, 2005). Neste caso, a taxa leva em consideração não só o distrito em si, mas também os seus vizinhos. Apesar de essa suavização ter produzido resultados melhores do que para a taxa bruta, o coeficiente de silhueta ainda ficou muito próximo de 0.

Com a grande quantidade de LISA *Clusters* não significativos, entendemos que para uma aplicação multivariada, onde estão sendo analisadas várias doenças em conjunto com o intuito de compará-las, essa metodologia não produziu os resultados esperados.

5.3 Agrupamento por doença

O agrupamento por doença busca responder a primeira questão de pesquisa definida: "Quais grupos de doenças causadoras de morte na cidade de São Paulo possuem distribuições geográficas similares?". Ou seja, estamos buscando um agrupamento das doenças de acordo com os comportamentos das taxas de mortalidade nos distritos administrativos.

Conforme descrito no Capítulo 4, antes da aplicação do algoritmo de agrupamento hierárquico foi construído o dendrograma para que se pudesse definir o número ideal de *clusters*. Esse diagrama está apresentado na Figura 5.8 e nela é ilustrado o ponto de corte testado, com 4 *clusters*. Essa escolha foi feita pois visualmente é a que apresenta maiores ganhos, e também ao testar um maior número de *clusters* o coeficiente de silhueta cai significativamente. O gráfico mostra que, com o ponto de corte estabelecido, a maioria dos capítulos da CID 10 estão distribuídos em um grande *cluster* e três capítulos ficam isolados, pertencendo cada um a um *cluster* distinto.

Após a aplicação do algoritmo, podemos ver exatamente essa separação dos 4 grupos. A divisão dos capítulos dentro dos *clusters* está listada na Tabela 5.3. Para uma melhor interpretação dos grupos criados, foram calculadas as médias simples das taxas de mortalidade padronizadas de todas as doenças pertencentes a um mesmo *cluster* e foi criado um mapa da distribuição geográfica dessas taxas para cada um desses 4 grupos.

Os capítulos I, III, IV, V, VI, XI, XII, XIII, XIV, XVIII e XX foram todos alocados em um único grupo, o *cluster* 1. As taxas médias de mortalidade para esse grupo podem ser observadas na Figura 5.9a. Para esse mapa, é importante destacar que a escala de valores é toda negativa. Isso significa que esses capítulos, no geral, possuem taxas abaixo das médias gerais para todos os distritos, ou seja, são capítulos que possuem baixas taxas de mortalidade em comparação aos demais em todos os distritos administrativos. Ainda nesse mapa, vemos alguns distritos em destaque, dentre eles Marsilac. Como apontado na Seção 5.1, esse distrito possui uma população baixa em comparação aos demais e pode gerar alguns *outliers*, pois a taxa de mortalidade se torna muito volátil. Outros distritos que se destacam nesse mapa são Brás, Pari, Tremembé e Vila Andrade, por possuírem taxas um pouco acima das demais regiões, mas ainda com a ressalva que são taxas abaixo das médias desses distritos e também com o ponto de atenção de que a variação da escala de valores do mapa é bastante baixa (de -0,300 a -0,265, uma variabilidade de 12%). Nesse

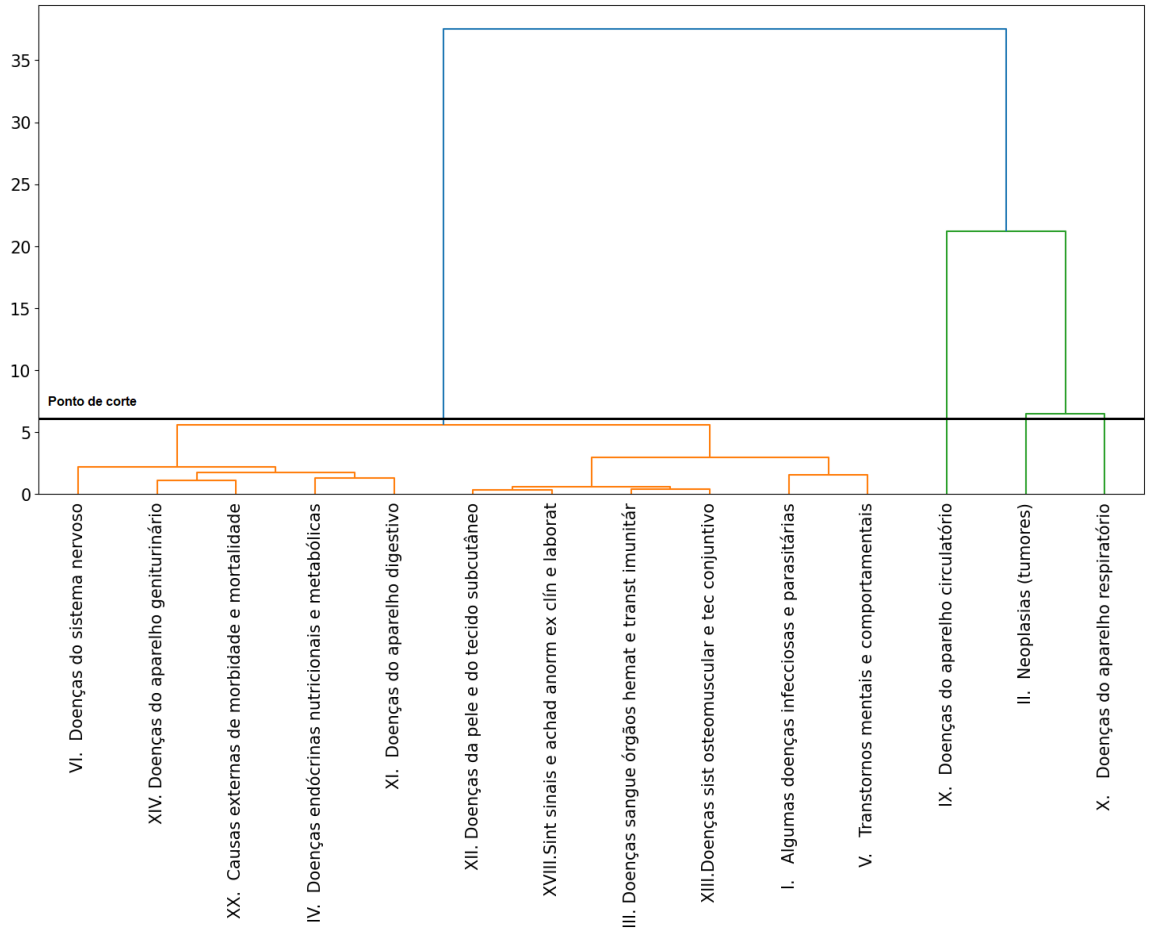


Figura 5.8: *Dendrograma construído para o agrupamento de doenças.*

Capítulo CID 10	*Cluster*
I. Algumas doenças infecciosas e parasitárias III. Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários IV. Doenças endócrinas, nutricionais e metabólicas V. Transtornos mentais e comportamentais VI. Doenças do sistema nervoso XI. Doenças do aparelho digestivo XII. Doenças da pele e do tecido subcutâneo XIII. Doenças do sistema osteomuscular e do tecido conjuntivo XIV. Doenças do aparelho geniturinário XVIII. Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte XX. Causas externas de morbidade e mortalidade	1
IX. Doenças do aparelho circulatório	2
X. Doenças do aparelho respiratório	3
II. Neoplasias [tumores]	4

Tabela 5.3: *Agrupamento de doenças com 3 clusters.*

grupo, não existe uma região específica de destaque para essas doenças, mas sim alguns distritos isolados com comportamentos que se destacam superficialmente.

O capítulo IX, doenças do aparelho circulatório, foi alocado no *cluster* 2. As taxas médias de mortalidade para esse grupo podem ser observadas na Figura 5.9b. Diferente do *cluster* 1, esse mapa possui a escala de valores positiva. Ou seja, é um capítulo que possui altas taxas de mortalidade em comparação aos demais em todos os distritos administrativos. Neste caso, podemos observar que as regiões de destaque estão localizadas na periferia da cidade de São Paulo. Ou seja, esse capítulo possui taxas de mortalidade acima das médias para todos os distritos, porém na periferia esse comportamento é ainda mais evidente.

O capítulo X, doenças do aparelho respiratório, ficou alocado no *cluster* 3 e o capítulo II, neoplasias, no *cluster* 4. Os mapas desses novos grupos podem ser observados nas Figuras 5.9c e 5.9d, respectivamente. Neles, podemos ver que ambos possuem a escala de valores positiva, porém o *cluster* 4 possui valores mais elevados. Como visto na Seção 5.1, o capítulo II é um dos que possuem os maiores percentuais de óbitos, então faz sentido a escala de valores desse mapa ser mais elevada que para o capítulo X. Para o *cluster* 3, o mapa está mais homogêneo e o distrito Marsilac fica em evidência, possivelmente devido à ocorrência de valores de baixa significância estatística. Já para o *cluster* 4, apesar de todos os distritos terem altas taxas de mortalidade, a região central tem um destaque maior.

Por fim, foi utilizado o coeficiente de silhueta para avaliar a adequação do modelo criado. O coeficiente calculado foi de 0,62. Como o coeficiente varia entre -1 e 1, consideramos que o modelo está bem ajustado.

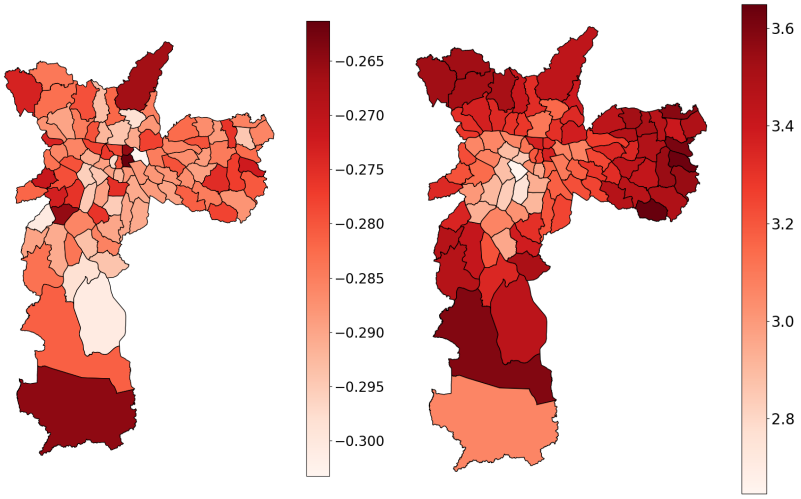
5.3.1 Discussão

Apesar de o *cluster* 1 contemplar capítulos com taxas de mortalidade abaixo das médias para todos os distritos administrativos, é importante trazer a visão de que existem algumas regiões que necessitam de uma atenção maior em relação a essas causas. Esse grupo contempla 11 dos 14 capítulos da CID 10 analisados e podem ser trabalhados isoladamente em trabalhos futuros, para que se possa entender um pouco melhor do seu comportamento e identificar se há diferenças entre eles, que não foram apontadas quando analisados todos os capítulos em conjunto.

O capítulo IX, doenças do aparelho circulatório, foi identificado como tendo um comportamento específico. Esse é um grupo de doenças que causa mortes em toda a cidade de São Paulo, mas ele possui um agravante maior na região periférica, que coincide com a região com menores índices de desenvolvimento humano.

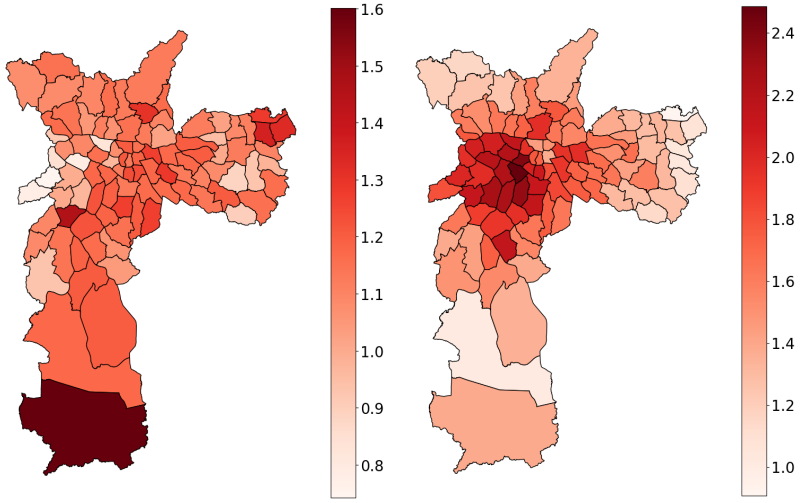
Para o capítulo X, doenças do aparelho respiratório no *cluster* 3, vemos que existe uma homogeneidade na distribuição geográfica dessa doença, exceto pelo *outlier* em Marsilac. Esse comportamento não desperta nenhum ponto de atenção de imediato, porém uma análise específica para esse capítulo dos códigos da CID 10 que o contemplam pode ser realizada em trabalhos futuros para entender se alguma causa em específico pode ter um comportamento diferente.

Já para o capítulo II, neoplasias, vemos um comportamento inverso do identificado para o capítulo IX. Neste caso, as regiões central e oeste possuem maiores destaques



(a) *Cluster 1.*

(b) *Cluster 2.*



(c) *Cluster 3.*

(d) *Cluster 4.*

Figura 5.9: *Mapa das distribuições geográficas das taxas médias de mortalidade dos capítulos por cluster.*

em relação às taxas de mortalidade, áreas que coincidem com as de maiores índices de desenvolvimento humano. A população mais rica tende a viver mais e, então, morrer de câncer.

Em resumo, os resultados mostraram que morrem mais pessoas de doenças do aparelho circulatório nas áreas mais vulneráveis da cidade, enquanto morrem mais pessoas de neoplasias nas regiões com maiores condições socioeconômicas. Como as neoplasias são doenças que ainda não se sabe a cura, este estudo mostra que as pessoas com melhores condições de vida morrem dessas causas pois para as demais, conseguem boas condições de tratamento. Enquanto isso, a população mais desamparada da cidade morre de causas que têm cura e prevenção em muitos casos, pois falta acesso a tratamentos de saúde com qualidade.

Esse primeiro modelo deixa claro que, apesar de existirem causas de morte que abrangem toda a cidade, existem regiões com uma maior vulnerabilidade e, consequentemente, requerem uma atenção especial. Portanto, a criação de políticas públicas específicas para cada um dos capítulos existentes podem começar a ser desenhadas a partir dessas regiões destacadas.

5.4 Agrupamento por distrito

O agrupamento por distrito busca responder a segunda questão de pesquisa definida: "Quais distritos administrativos da cidade de São Paulo possuem comportamentos similares em relação às doenças causadoras de morte?". Ou seja, estamos buscando um agrupamento dos distritos administrativos de acordo com os comportamentos das taxas de mortalidade nas doenças estudadas.

Assim como para o agrupamento anterior, o dendrograma foi construído (Figura 5.10) e o ponto de corte definido criou um agrupamento dos distritos administrativos em 4 *clusters*.

Como agora estamos agrupando distritos administrativos, conseguimos visualizar através do mapa da Figura 5.11 a distribuição dos *clusters*. Nele, podemos ver que o *cluster* 1 é composto por uma parte da região central e uma parte da região oeste, enquanto o *cluster* 2 contempla praticamente os arredores do *cluster* 1. O *cluster* 3 é composto apenas pelo distrito Marsilac e o último grupo contempla os distritos das regiões mais periféricas da cidade.

Para uma melhor interpretação dos grupos identificados, calculamos as médias simples das taxas de mortalidade padronizadas de todos os distritos administrativos pertencentes a um mesmo *cluster*, para cada doença (Figura 5.12). A partir desse gráfico, podemos obter conclusões semelhantes àquelas obtidas na Seção 5.3. Ou seja, os capítulos II, IX e X são os que possuem as maiores taxas médias, em todos os *clusters* e os demais capítulos possuem taxas médias abaixo de 0, com algumas exceções pontuais.

Para o capítulo II, neoplasias, vemos que as taxas médias dos *clusters* 1 e 2 se sobressaem em relação às taxas médias dos *clusters* 3 e 4. Isso mostra que, apesar desse capítulo ter taxas de mortalidade acima da média dos demais em todos os distritos administrativos, essas taxas possuem um maior destaque nas regiões dos dois primeiros grupos identificados,

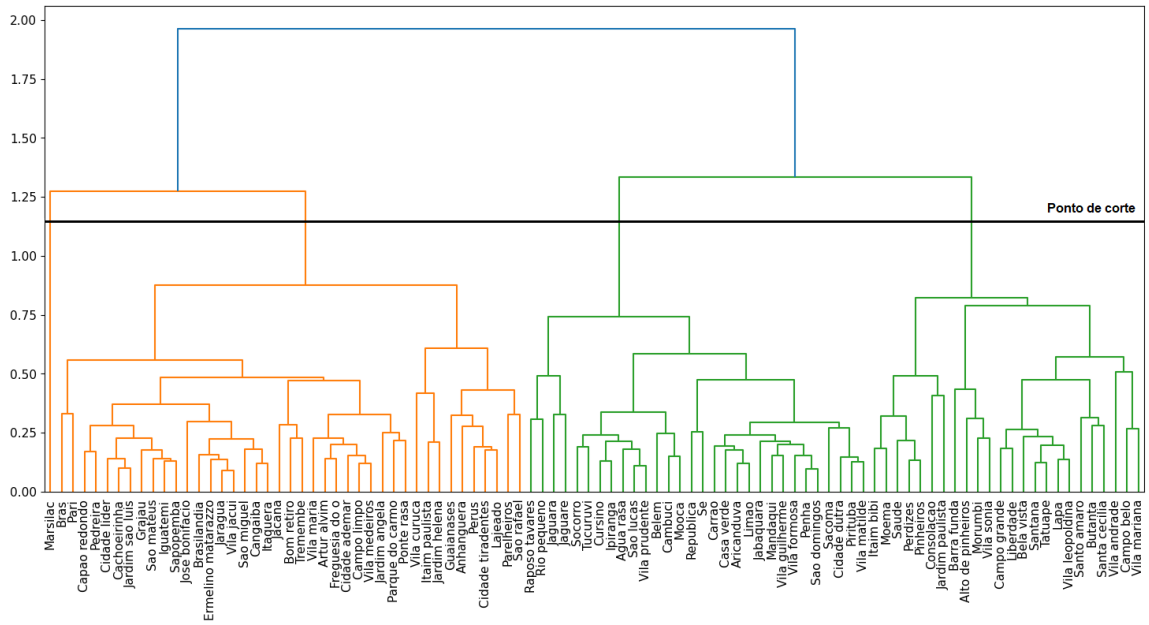


Figura 5.10: *Dendrograma construído para o agrupamento de distritos administrativos.*

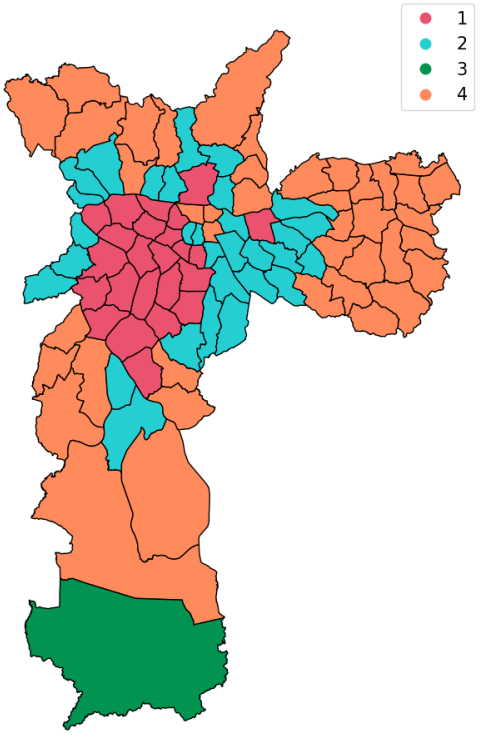


Figura 5.11: *Distribuição geográfica dos clusters.*

que correspondem à região central da cidade de São Paulo e seus arredores. Para o capítulo IX, doenças do aparelho circulatório, as taxas médias dos *clusters* 2 e 4 estão em destaque em relação aos outros dois grupos, ou seja, a área periférica da cidade. Já para os capítulos X, doenças do aparelho respiratório, e XX, causas externas de morbidade e mortalidade, o *cluster* 3 está em destaque em relação aos demais, que é o distrito de Marsilac. Para ambos os capítulos, Marsilac possui uma taxa média maior que para os demais distritos e, exclusivamente para o capítulo XX, todos os demais distritos possuem taxas abaixo da média. Porém, como já destacadas as considerações sobre Marsilac, podemos considerar essas conclusões pouco fidedignas.

Em relação ao capitulo IV, doenças endócrinas, nutricionais e metabólicas, as taxas médias estão abaixo de 0 para todos os grupos, porém vale destacar que para os *clusters* 1 e 2 as taxas são menores que para os demais, indicando que na região central essas doenças se destacam por terem taxas menores que as demais. Apesar de todas as taxas para esse capítulo serem muito próximas de 0, essa conclusão pode trazer uma informação relevante. Um outro ponto interessante é sobre o capítulo VI, doenças do sistema nervoso, que possui taxas médias abaixo de 0 para todos os *clusters*, exceto para o primeiro. Isso indica que os óbitos causados por esse capítulo estão abaixo da média para todos os distritos, exceto para aqueles concentrados em uma parcela da região central e na região oeste.

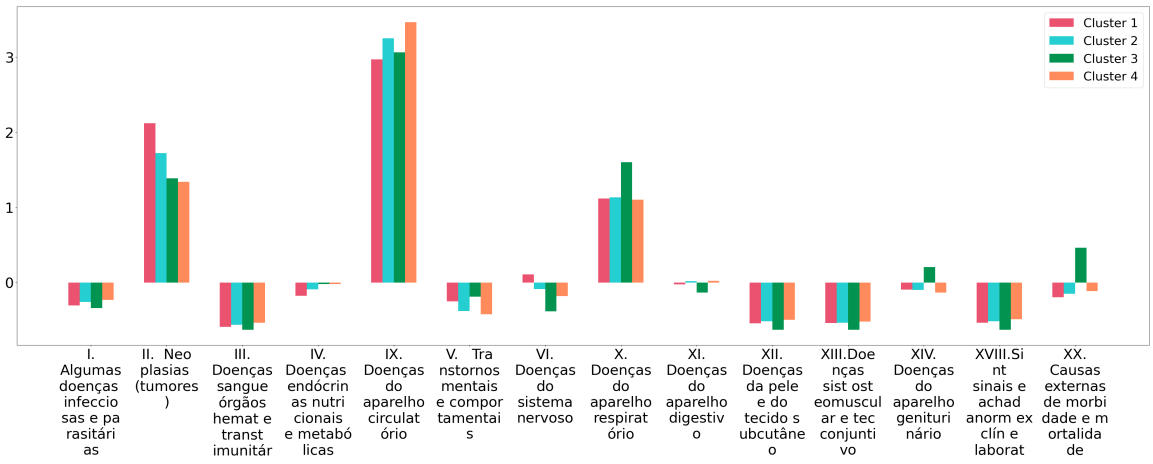


Figura 5.12: *Médias das taxas de mortalidade padronizadas por distrito administrativo, para cada cluster criado.*

Para esse teste, o coeficiente de silhueta calculado foi de 0,70. Ou seja, um ajuste muito bom também e melhor que para o agrupamento de doenças.

5.4.1 Discussão

As conclusões obtidas no agrupamento por distrito administrativo foram muito similares às conclusões obtidas no agrupamento por doenças, que era o principal objetivo ao realizarmos a análise sob essas duas perspectivas.

Para o capítulo II, neoplasias, obtivemos a mesma conclusão da Seção 5.3 de que há uma maior incidência de óbitos com essas causas na região central da cidade, porém neste caso podemos considerar uma maior área de abrangência quando nos referimos à região

central, do que quando observamos o mapa da Figura 5.9d. Já para os capítulo IX, doenças do aparelho circulatório, e X, doenças do aparelho respiratório, as áreas de abrangência identificadas são muito similares àquelas observadas na seção anterior. Além disso, nesse formato o algoritmo foi capaz de identificar o distrito de Marsilac como um *outlier* e ele ficou isolado em um *cluster*.

Essa abordagem por distrito trouxe algumas conclusões novas que não foram identificadas na abordagem por doença pois muitos capítulos foram alocados em um único *cluster*. Apesar de as taxas médias para o capítulo IV, doenças endócrinas, nutricionais e metabólicas, serem muito próximas de 0, a visão de que essa taxa é ainda menor para uma parcela das regiões central e oeste também pode ser correlacionada com a condição socioeconômica desses locais. Essas são doenças que podem ser prevenidas com uma boa alimentação e com a prática de atividades físicas, que está mais acessível apenas para a população de alta renda. Essas mesmas regiões do *cluster* 1 também se destacam, porém de uma forma negativa, para os óbitos causados pelo capítulo VI, doenças do sistema nervoso. Nesse caso, as taxas de mortalidade estão acima da média para uma parcela das regiões central e oeste, e abaixo da média para as demais localidades. Estudos futuros também podem ajudar a entender esse comportamento e suportar políticas de melhoria para essa população.

No geral, os resultados obtidos com a abordagem por distrito administrativo possuem ganhos em relação à abordagem por doença, pois conseguimos obter novas conclusões relevantes e, além disso, essa perspectiva resultou em um coeficiente de silhueta maior.

O próximo capítulo resume as conclusões obtidas neste capítulo, e também aborda quais foram as lições aprendidas durante nossa pesquisa e quais trabalhos futuros podem ser desenvolvidos a partir dos resultados apresentados.

Capítulo 6

Conclusões

No Capítulo 1 foram definidas as seguintes questões de pesquisa:

1. Quais grupos de doenças causadoras de morte na cidade de São Paulo possuem distribuições geográficas similares?
2. Quais distritos administrativos da cidade de São Paulo possuem comportamentos similares em relação às doenças causadoras de morte?

As duas questões possuem objetivos complementares, com o intuito de obtermos resultados sob duas perspectivas diferentes e isso se concretizou. Os resultados apresentados no Capítulo 5 para ambas abordagens são bastante similares, ou seja, as conclusões gerais são as mesmas e serão detalhadas a seguir. O código fonte utilizado para a obtenção de todos os resultados apresentados está disponível em repositórios do GitLab[1].

No geral, foi identificado que dois capítulos da CID 10 possuem distribuições geográficas específicas. Apesar de as doenças do capítulo II, neoplasias, apresentar uma alta taxa de mortalidade para todo o município, quando calculadas as taxas padronizadas para cada distrito administrativo vemos que há uma concentração maior dessa taxa na região central da cidade. Isso mostra que as áreas com maiores condições socioeconômicas são mais problemáticas para doenças que não se sabe a cura, pois para as demais causas essa população está amparada. Enquanto isso, a população com menos condições socioeconômicas acaba indo a óbito antes, por outras causas, por possuírem menor acesso a um atendimento de saúde de qualidade. Já para o capítulo IX, doenças no sistema circulatório, o comportamento é exatamente o oposto, ou seja, as maiores taxas de mortalidade estão na região periférica da cidade, uma vez que essa população está mais desprotegida em relação ao acesso à saúde.

Para o capítulo IV, doenças endócrinas, nutricionais e metabólicas, vimos que uma área formada por uma parcela da região central e uma parcela da região oeste possui taxas de mortalidade menores que para as demais regiões. Essas são doenças que são prevenidas com boa alimentação e prática de atividades físicas, o que só a população de mais alta renda consegue ter. Por fim, para as doenças do capítulo VI, doenças do sistema nervoso, vemos

[1] https://gitlab.com/interscity/health/geospatial-analysis-of-deaths/-/tree/main/

um comportamento oposto. Essa área destacada por uma parcela da região central e uma parcela da região oeste possui taxas maiores dessa causa que as demais áreas. Portanto, entender o porquê isso ocorre pode trazer melhores percepções para o desenvolvimento em políticas públicas de saúde. Para os demais capítulos da CID 10, concluiu-se que não existe uma distribuição geográfica específica.

A Seção 6.1 apresenta as lições aprendidas durante a execução das análises apresentadas e a Seção 6.2 sugere alguns trabalhos que podem ser desenvolvidos no futuro em continuação a essa pesquisa.

6.1 Lições aprendidas

No Capítulo 3 vimos que existem diversas formas de realizar análises geoespaciais, e nem todas foram testadas nesta pesquisa. Além disso, todas as análises listadas na Seção 5.2 devem servir de aprendizado sobre o que já foi testado e não produziu os resultados esperados.

Quando trabalhamos com taxas de mortalidade, é sempre importante levar-se em consideração as distribuições de sexo e faixa etária da população estudada. Para entender essa necessidade, construir análises descritivas antes de iniciar as análises principais é muito importante e essas análises podem mudar o rumo da pesquisa. Além disso, trabalhar com capítulos ao invés dos códigos específicos das doenças reduz as conclusões enviesadas por conta de *outliers*. Por outro lado, acabamos obtendo conclusões que abrangem um conjunto de muitas doenças e, se alguma causa específica possui um comportamento estatisticamente significativo e diferente do comportamento geral do capítulo, isso não é evidenciado nessa abordagem.

O Índice de Moran é uma técnica de análise espacial que foi extensivamente testada durante esta pesquisa de mestrado. Embora essa seja uma técnica que funciona muito bem quando estamos lidando com doenças específicas, para o caso multivariado ela não trouxe os resultados esperados, pois muitos distritos administrativos ficam classificados como não significativos e o algoritmo de agrupamento acaba não sendo efetivo. Portanto, em trabalhos futuros pode ser interessante testar outras técnicas de análise espacial ao invés do Índice de Moran.

De modo geral, aprendeu-se que não é trivial definir uma metodologia para o objetivo dessa pesquisa e, além disso, não é trivial determinar como os dados devem ser construídos. Quais padronizações devem ser realizadas, suavizações, agrupamentos, etc. Portanto, entende-se que esta pesquisa trouxe muitos aprendizados, mas ainda há muito para ser explorado nesse contexto.

6.2 Trabalhos futuros

Existem diversos trabalhos que podem ser executados com base nesta pesquisa. Uma análise de correlação entre os grupos identificados com variáveis externas, como geografia física da região de análise, fatores meteorológicos, geografias humana, social, política e econômica, pode trazer contribuições significativas para a elaboração de políticas públicas.

Esse tipo de análise pode auxiliar no entendimento do comportamento das taxas de mortalidade para o capítulo VI, como mencionado anteriormente. Relacionar os dados extraídos com dados de internações ou autópsias também pode trazer informações muito relevantes, visto que muitas vezes o paciente é internado com uma causa, mas a certidão de óbito consta uma causa diferente da internação.

Além disso, pode ser interessante a aplicação da metodologia para os capítulos da CID 10 sem considerar aqueles para os quais já obtivemos conclusões (ou seja, sem os capítulos II, IV, VI, IX e X), para que se possa identificar distribuições geográficas específicas para aqueles capítulos que acabaram ficando todos agrupados em um único *cluster*.

Em relação ao mapa coroplético construído nas análises preliminares (Figura 4.5), pode ser bastante proveitoso utilizá-lo como um coeficiente de heterogeneidade na cidade, afim de medir a desigualdade da região. Assim, poderiam ser construídos mapas com este coeficiente para diferentes doenças e então auxiliar na identificação de políticas públicas virtuosas.

Um ponto importante é que uma das motivações desta pesquisa foi criar uma metodologia que pudesse ser aplicada em outros casos, então um outro trabalho futuro seria aplicar essa metodologia para dados mais granulares que capítulos e menos granulares que os próprios códigos da CID 10, para estudar se novas conclusões podem ser obtidas. Até mesmo outros tipos de agrupamentos de doenças podem ser analisados, como por exemplo por doenças sensíveis ou não à atenção primária, sensíveis ou não à poluição, sensíveis ou não a exercícios físicos, etc. Outras granularidades de regiões de análise também podem ser exploradas, como subdistritos ao invés de distritos. Inclusive, dados subdistritais podem contribuir significativamente com políticas de saúde de precisão. Além disso, a metodologia pode ser explorada para outras regiões de análise além da cidade de São Paulo, como por exemplo o país inteiro sendo a granularidade de regiões analisadas os municípios brasileiros, ou estados.

Entende-se também que há muito valor na metodologia desenvolvida como uma ferramenta genérica, ou seja, ela pode ser aplicada em contextos distintos, como por exemplo na análise espacial da taxa de vacinação de acordo com os diversos tipos de vacina disponíveis na rede pública. Esse tipo de análise possibilitaria a criação de políticas voltadas a campanhas de vacinação específicas. Um outro exemplo de aplicação é na análise espacial de boletins de ocorrência de acordo com o tipo da ocorrência. Neste caso, seria possível criar políticas específicas para cada tipo de ocorrência nas regiões de maior necessidade.

Assim, espera-se que tanto pesquisadores quanto gestores de saúde possam usufruir dos resultados obtidos em nossa pesquisa, e que as conclusões aqui obtidas possam evoluir com o surgimento de novos trabalhos.

Apêndice A

Distritos administrativos

O município de São Paulo é adminsitrativamente dividido em 96 distritos, conforme indicado na Figura A.1 obtida através do site *Encontra São Paulo* (EncontraSãoPaulo, 2008). Essa divisão territorial é mais agrupada que o nível de bairros e mais segmentada que o nível de subprefeituras.

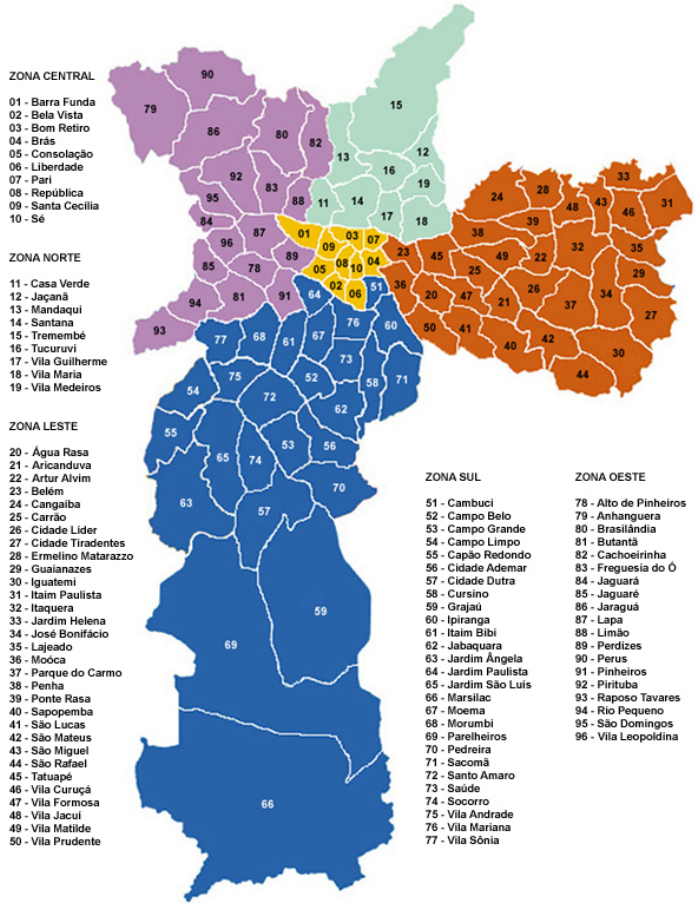


Figura A.1: *Mapa dos Distritos Administrativos da cidade de São Paulo.*

Anexo A

Números gerais por distrito administrativo

Alguns números gerais por distrito administrativo da cidade de São Paulo, como as suas áreas, estimativas populacionais médias e médias de óbitos por ano entre 2014 e 2018 estão disponíveis nas Tabelas A.1 a A.3.

Distrito administrativo	Área (km²)	Estimativa populacional (total)	Óbitos (total)	Estimativa populacional (40+)	Óbitos (40+)
Água Rasa	6,9	83.669	871	41.652	829
Alto de Pinheiros	7,7	41.877	314	23.207	308
Anhanguera	33,3	77.842	246	25.591	195
Aricanduva	6,6	87.154	710	38.572	658
Artur Alvim	6,6	102.179	779	43.463	705
Barra Funda	5,6	15.384	158	7.162	148
Bela Vista	2,6	71.963	492	33.174	459
Belém	6,0	47.747	413	20.267	379
Bom Retiro	4,0	37.149	224	14.269	198
Brás	3,5	31.700	235	11.912	202
Brasilândia	21,0	275.188	1.656	92.540	1.358
Butantã	12,5	54.087	479	26.275	457
Cachoeirinha	13,3	145.007	1.004	52.405	851
Cambuci	3,9	39.408	310	17.955	291
Campo Belo	8,8	64.607	626	33.443	597
Campo Grande	13,1	104.696	671	48.631	628
Campo Limpo	12,8	222.251	1.107	79.324	920
Cangaíba	16,0	137.371	962	55.062	860
Capão Redondo	13,6	285.539	1.310	97.464	1.066

Tabela A.1: *Informações de área, estimativa populacional média e média de óbitos por ano dos distritos administrativos de São Paulo (1/3).*

Distrito administrativo	Área (km²)	Estimativa populacional (total)	Óbitos (total)	Estimativa populacional (40+)	Óbitos (40+)
Carrão	7,5	84.509	789	41.173	748
Casa Verde	7,1	85.991	771	37.955	718
Cidade Ademar	12,0	278.705	1.596	101.117	1.384
Cidade Dutra	29,3	199.849	1.175	76.363	1.043
Cidade Líder	10,2	131.948	810	50.144	709
Cidade Tiradentes	15,0	225.693	1.048	71.048	827
Consolação	3,7	57.454	440	29.071	421
Cursino	12,8	112.157	870	52.313	816
Ermelino Matarazzo	8,7	116.896	690	44.061	603
Freguesia do Ó	10,5	141.025	1.196	61.671	1.096
Grajaú	92,0	378.357	1.786	122.892	1.423
Guaianases	8,6	107.196	622	36.275	521
Iguatemi	19,6	140.984	682	45.994	548
Ipiranga	10,5	110.279	890	50.940	837
Itaim Bibi	9,9	95.832	711	49.969	690
Itaim Paulista	12,0	230.350	1.209	78.671	1.025
Itaquera	14,6	208.764	1.313	78.602	1.141
Jabaquara	14,1	227.188	1.543	95.591	1.413
Jaçanã	7,8	95.553	679	37.717	605
Jaguara	4,6	24.313	214	11.227	201
Jaguaré	6,6	53.207	279	20.804	248
Jaraguá	27,6	202.446	1.056	70.309	884
Jardim Ângela	37,4	321.607	1.322	101.359	1.030
Jardim Helena	9,1	134.680	821	45.116	694
Jardim Paulista	6,1	90.156	679	48.194	659
Jardim São Luís	24,7	283.821	1.421	100.313	1.184
José Bonifácio	14,1	132.128	704	49.857	608
Lajeado	9,2	170.298	953	55.445	779
Lapa	10,0	66.913	661	35.048	639
Liberdade	3,7	71.582	503	33.580	472
Limão	6,3	79.512	680	33.130	617
Mandaqui	13,1	108.797	811	49.752	758
Marsilac	200,0	8.320	38	2.996	31
Moema	9,0	87.574	567	47.084	551
Mooca	7,7	78.938	774	38.977	745
Morumbi	11,4	50.875	291	23.178	266
Parelheiros	153,5	144.707	722	46.091	593
Pari	2,9	18.429	152	7.398	135
Parque do Carmo	15,4	70.415	440	25.739	370
Pedreira	18,7	154.427	658	53.317	561

Tabela A.2: *Informações de área, estimativa populacional média e média de óbitos por ano dos distritos administrativos de São Paulo (2/3).*

Distrito administrativo	Área (km²)	Estimativa populacional (total)	Óbitos (total)	Estimativa populacional (40+)	Óbitos (40+)
Penha	11,3	128.822	1.117	59.478	1.039
Perdizes	6,1	113.778	811	58.524	783
Perus	23,9	85.660	456	28.981	382
Pinheiros	8,0	65.832	507	35.048	493
Pirituba	17,1	170.028	1.229	71.721	1.114
Ponte Rasa	6,4	91.269	684	38.962	619
Raposo Tavares	12,6	104.653	533	38.146	464
República	2,3	60.358	460	27.584	406
Rio Pequeno	9,7	121.791	748	48.253	678
Sacomã	14,2	257.849	1.493	104.300	1.349
Santa Cecília	3,9	87.062	661	41.851	619
Santana	12,6	115.185	1.093	58.276	1.042
Santo Amaro	15,6	73.767	661	38.635	635
São Domingos	10,0	85.809	597	36.265	546
São Lucas	9,9	142.923	1.087	64.258	1.005
São Mateus	13,2	155.007	1.033	60.044	902
São Miguel	7,5	89.967	778	34.959	695
São Rafael	13,0	153.562	683	52.847	556
Sapopemba	13,5	287.803	1.718	108.921	1.487
Saúde	8,9	133.258	987	69.115	957
Sé	2,1	25.606	170	9.535	146
Socorro	12,9	36.684	323	17.907	305
Tatuapé	8,2	94.636	819	47.909	784
Tremembé	56,3	213.598	1.283	80.029	1.118
Tucuruvi	9,0	97.333	914	47.625	874
Vila Andrade	10,3	150.324	494	51.551	410
Vila Curuçá	9,7	151.394	908	54.014	776
Vila Formosa	7,4	94.493	808	44.467	758
Vila Guilherme	6,9	56.124	483	25.412	453
Vila Jacuí	7,7	144.265	845	51.764	718
Vila Leopoldina	7,2	43.262	236	19.733	224
Vila Maria	11,8	113.651	810	44.061	715
Vila Mariana	8,6	131.765	1.088	68.504	1.041
Vila Matilde	8,9	105.445	838	47.400	784
Vila Medeiros	7,7	125.595	1.071	52.891	977
Vila Prudente	9,9	104.590	843	48.640	791
Vila Sônia	9,9	116.994	648	48.431	593

Tabela A.3: *Informações de área, estimativa populacional média e média de óbitos por ano dos distritos administrativos de São Paulo (3/3).*

Referências

[Almeida 2018] Luiz Gustavo de Almeida. *John Snow na Guerra das Epidemias*. 2018. url: https://www.revistaquestaodeciencia.com.br/questao-nerd/2019/04/15/john-snow-na-guerra-das-epidemias (acesso em 24/12/2021) (citado nas pgs. 17, 18).

[Aprilliant 2021] Audhi Aprilliant. *The k-modes as Clustering Algorithm for Categorical Data Type*. 2021. url: https://medium.com/geekculture/the-k-modes-as-clustering-algorithm-for-categorical-data-type-bcde8f95efd7 (acesso em 22/02/2023) (citado nas pgs. 10, 11).

[Barata 2013] Rita Barradas Barata. "Epidemiologia e políticas públicas". Em: *Revista Brasileira de Epidemiologia* 16 (2013), pgs. 3–17 (citado na pg. 2).

[Barcellos *et al.* 2005] Christovam Barcellos, Adelaide Kreutz Pustai, Maria Angélica Weber e Maria Regina Varnieri Brito. "Identificação de locais com potencial de transmissão de dengue em porto alegre através de técnicas de geoprocessamento". Em: *Revista da Sociedade Brasileira de Medicina Tropical* 38 (2005), pgs. 246–250 (citado nas pgs. 2, 18, 19).

[Canada 2017] Statistics Canada. *Age-standardized Rates*. 2017. url: https://www.statcan.gc.ca/en/dai/btd/asr (acesso em 11/12/2022) (citado na pg. 6).

[CID10 2013] CID10. *CID10*. 2013. url: https://cid10.com.br/ (acesso em 23/01/2022) (citado nas pgs. 2, 25).

[Costa-Nobre *et al.* 2021] Daniela Testoni Costa-Nobre *et al.* "Clusters of cause specific neonatal mortality and its association with per capita gross domestic product: a structured spatial analytical approach". Em: *PLOS ONE* 16 (2021), pgs. 1–12 (citado nas pgs. 19, 43, 45).

[Czeresnia e Ribeiro 1997] Dina Czeresnia e Adriana Maria Ribeiro. "O conceito de espaço em epidemiologia: uma interpretação histórica e epistemológica". Em: *Cadernos de Saúde Pública* 13 (1997), pgs. 585–593 (citado na pg. 1).

[Dados Abertos 2015] Guia de Dados Abertos. *Shapefile*. 2015. url: https://ceweb.br/guias/dados-abertos/capitulo-41/ (acesso em 17/12/2022) (citado na pg. 27).

[Druck *et al.* 2004] Suzana Druck, Marília Sá Carvalho, Gilberto Câmara e Antônio Miguel Vieira Monteiro. *Análise Espacial de Dados Geográficos*. EMBRAPA, 2004 (citado nas pgs. 1, 13).

[EncontraSãoPaulo 2008] EncontraSãoPaulo. *Mapa da Cidade de São Paulo*. 2008. url: https://www.encontrasaopaulo.com.br/mapa-de-sao-paulo.php (acesso em 15/01/2022) (citado na pg. 57).

[Figueiredo *et al.* 2001] Cláudia Maria de Figueiredo *et al.* "Leptospirose humana no município de belo horizonte,minas gerais, brasil: uma abordagem geográfica". Em: *Revista da Sociedade Brasileira de Medicina Tropical* 34 (2001), pgs. 331–338 (citado nas pgs. 2, 17).

[Geografia e Estatística 2017] Instituto Brasileiro de Geografia e Estatística. *IBGE Cidades*. 2017. url: https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama (acesso em 05/02/2023) (citado na pg. 37).

[Hino *et al.* 2011] Paula Hino, Tereza Cristina Scatena Villa, Tarcísio Neves da Cunha e Claudia Benedita dos Santos. "Distribuição espacial de doenças endêmicasno município de ribeirão preto (sp)". Em: *Ciência & Saúde Coletiva* 16 (2011), pgs. 1289–1294 (citado nas pgs. 1, 3, 18).

[J.Rousseeuw 1987] Peter J.Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". Em: *Journal of Computational and Applied Mathematics* 20 (1987), pgs. 53–65 (citado nas pgs. 12, 13).

[Jayasekara *et al.* 2013] J M K B Jayasekara, D M Dissanayake, S B Adhikari e P Bandara. "Geographical distribution of chronic kidney disease of unknown origin in north central region of sri lanka". Em: *Ceylon Medical Journal* 58 (2013) (citado nas pgs. 18, 19).

[KELLY-HOPE *et al.* 2007] LOUISE A. KELLY-HOPE *et al.* "Geographical distribution and risk factors associated with enteric diseases in vietnam". Em: *The American Society of Tropical Medicine and Hygiene* 76 (2007), pgs. 706–712 (citado nas pgs. 19, 20).

[Ludermir 2021] Teresa Bernarda Ludermir. "Inteligência artificial e aprendizado de máquina: estado atual e tendências". Em: *Estudos Avançados* 35 (2021), pgs. 85–94 (citado na pg. 5).

[Luzardo *et al.* 2017] Antonio José Rocha Luzardo, Rafael March Castañeda Filho e Igor Brum Rubim. "Análise espacial exploratória com o emprego do índice de moran". Em: *GEOgraphia* 19 (2017), pgs. 161–179 (citado nas pgs. 13, 15).

[Maklin 2018] Cory Maklin. *Hierarchical Agglomerative Clustering Algorithm*. 2018. url: https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019 (acesso em 11/12/2022) (citado nas pgs. 8, 9).

[Mendonça *et al.* 2014] Francisco Mendonça, Wiviany Mattozo de Araújo e Thiago Kich Fogaça. "A geografia da saúde no brasil: estado da arte e alguns desafios". Em: *Investigaciones Geográficas* 48 (2014), pgs. 41–52 (citado na pg. 2).

[Oliveira 2021] Paulo Marcelo Rayner Oliveira. *Agrupamento Hierárquico (Hierarchical clustering)*. 2021. url: https://edu.taugc.com/blog/agrupamento-hierarquico-hierarchical-clustering/ (acesso em 11/12/2022) (citado na pg. 8).

[Opromolla *et al.* 2006] Paula A Opromolla, Ivete Dalben e Márcio Cardim. "Análise geoestatística de casos de hanseníase no estado de são paulo, 1991-2002". Em: *Revista de Saúde Pública* 40 (2006), pgs. 907–913 (citado na pg. 19).

[Pedregosa *et al.* 2011] F. Pedregosa *et al.* "Scikit-learn: machine learning in Python". Em: *Journal of Machine Learning Research* 12 (2011), pgs. 2825–2830 (citado na pg. 10).

[Pena *et al.* 2017] Marina Garcia Pena *et al.* "Clusterização espacial e não espacial: um estudo aplicado à agropecuária brasileira". Em: *TEMA (São Carlos)* 18.1 (2017) (citado nas pgs. 20, 21).

[Piech 2013] Chris Piech. *K Means*. 2013. url: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html (acesso em 09/01/2022) (citado nas pgs. 9, 11).

[Rodrigues-Júnior *et al.* 2008] Antonio Luiz Rodrigues-Júnior, Vinícius Tragante do Ó e Vivian Genaro Motti. "Estudo espacial e temporal da hanseníase no estado de são paulo, 2004-2006". Em: *Revista de Saúde Pública* 42 (2008), pgs. 1012–1020 (citado na pg. 19).

[Santos *et al.* 2005] Alexandre E. dos Santos, Alexandre L. Rodrigues e Danilo L. Lopes. "Aplicações de estimadores bayesianos empíricos para análise espacial de taxas de mortalidade". Em: *Simpósio Brasileiro de Geoinformática* 7 (2005), pgs. 303–309 (citado na pg. 45).

[Saúde 2023] Ministério da Saúde. *tabnet*. url: https://datasus.saude.gov.br/informacoes-de-saude-tabnet/ (acesso em 04/02/2023) (citado na pg. 26).

[Silva 2000] Luiz Jacintho da Silva. "O conceito de espaço na epidemiologia das doenças infecciosas". Em: *Cadernos de Saúde Pública* 16 (2000), pgs. 595–617 (citado na pg. 1).

[Snow 1999] John Snow. *Sobre a maneira de transmissão do cólera*. Hucitec, 1999 (citado nas pgs. 1, 17).

[Souza Dias *et al.* 2005] Márcia Célia Freitas de Souza Dias, Gutemberg Henrique Dias e Maurício Lisboa Nobre. "Distribuição espacial da hanseníase no município de mossoró/rn, utilizando o sistema de informação geográfica - sig". Em: *Anais Brasileiros de Dermatologia* 80 (2005), S289–S294 (citado nas pgs. 2, 18).

[Wang *et al.* 2012] Tao Wang, Fuzhong Xue, Yongjin Chen, Yunbo Ma e Yanxun Liu. "The spatial epidemiology of tuberculosis in linyi city, china, 2005–2010". Em: *BMC Public Health* 12.885 (2012) (citado nas pgs. 19, 45).