

UNIVERSIDADE FEDERAL DE VIÇOSA
CAMPUS DE RIO PARANAÍBA
SISTEMAS DE INFORMAÇÃO

RICARDO BORTOLUCCI PATRIANI DE CARVALHO

**ABORDAGENS COMPUTACIONAIS EXPLORATÓRIAS
PARA ESTUDO DO PERFIL EPIDEMIOLÓGICO DE
MUNICÍPIOS DE SÃO PAULO QUANTO À LEPTOSPIROSE**

RIO PARANAÍBA
2022

RICARDO BORTOLUCCI PATRIANI DE CARVALHO

ABORDAGENS COMPUTACIONAIS EXPLORATÓRIAS PARA
ESTUDO DO PERFIL EPIDEMIOLÓGICO DE MUNICÍPIOS DE SÃO
PAULO QUANTO À LEPTOSPIROSE

Monografia, apresentada ao curso de Sistemas de Informação da Universidade Federal de Viçosa - Campus Rio Paranaíba como requisito da obtenção do título de Bacharel em Sistemas de Informação

Orientador: Adriana Martinhago

Coorientador: Joelson Antônio dos Santos

RIO PARANAÍBA

2022

Resumo

Este trabalho consiste no estudo epidemiológico da Leptospirose no contexto do estado de São Paulo por meio de análises computacionais que consideram os indicadores sociais e o georreferenciamento. Essa doença foi eleita para o trabalho em questão por ser endêmica no Brasil e por acarretar em alto custo hospitalar para portadores, além de afastamento no trabalho e alta mortalidade. A Leptospirose é uma zoonose com ocorrência relacionada a situações precárias de saneamento básico e presença de roedores infectados por uma bactéria do gênero *Leptospira*. Segundo os dados do DATASUS, os municípios de São Paulo foram coletados como elementos multivvalorados que trazem indicadores de: urbanização, escolaridade e ocupação preenchidos com o número de casos equivalentes a cada contexto. Para complemento da base de dados foram coletados os dados de Densidade Populacional, IDH-M (IDH municipal) e de referenciamento geográfico dos municípios por meio do Instituto Brasileiro de Geografia e Estatística (IBGE). Ademais, a principal abordagem computacional utilizada foi a Mineração de Dados (MD) por meio da clusterização, utilizando: K-Médias (*K-Means*) e algoritmos de Clusterização Hierárquica, aplicadas em linguagem de programação *Python*, em que foram descobertos dois grupos de municípios com características convergentes. Desta forma, foram produzidas visualização em mapas por meio do Sistema de Gerenciamento de Banco de Dados PostgreSQL, com extensão geográfica, e por meio do Sistema de Informação Geográfica QGIS. Por meio do uso dessa segunda abordagem computacional foram geradas visualizações que ajudaram a levantar hipóteses sobre o conhecimento gerado pela MD.

Palavras-chave: Mineração de Dados, KDD, Leptospirose, Epidemiologia, Clusterização, Banco de Dados Geográficos.

Abstract

This paper consists of the epidemiological study of Leptospirosis in the context of the state of São Paulo through computational analysis that considers social indicators and georeferencing. This disease was chosen for this paper because it is endemic in Brazil and because it leads to high hospital costs for patients, as well as absence from work and high mortality. Leptospirosis is a zoonosis that is related to poor sanitation and the presence of rodents infected by a bacterium of the genus *Leptospira*. According to the availability of data in DATASUS, the municipalities of São Paulo were collected as multivalued elements that bring indicators of: urbanization, schooling and occupation filled with the number of cases equivalent to each context. To complement the database, data on Population Density, HDI-M (municipal HDI), and the geographic referencing of the municipalities through "Instituto Brasileiro de Geografia e Estatística" (IBGE) were collected. The computational approach was the data mining through the clustering approaches: K-Means and Hierarchical Clustering, in programming language Python, in which two groups of municipalities with convergent characteristics were discovered. From this result, a map visualization was produced by means of the PostgreSQL Database Management System, with geographic extension, and by means of the QGIS Geographic Information System. This second computational approach contributed to raise hypothesis according to knowledge provided by Data Mining.

Key-words: *Data Mining, KDD, Leptospirosis, Epidemiology, Clustering, Geo-database.*

Lista de ilustrações

Figura 1 – Processo do <i>KDD</i>	12
Figura 2 – Fluxo de Trabalho da Clusterização Hierárquica: Dendrograma	14
Figura 3 – Comparação Entre Representação Matricial e a Vetorial	19
Figura 4 – Combinações dos Algoritmos com os Números de Grupos (Ks) para Validação	25
Figura 5 – Subgrupo de Características extraídas do DATASUS e seus valores médios para cada agrupamento	27
Figura 6 – Subgrupo de Características extraídas do DATASUS e seus valores médios e desvio padrão para cada agrupamento	28
Figura 7 – Disposição Geográfica dos Grupos 0 e 1	29
Figura 8 – Disposição Geográfica do Grupo 0 com Graduação de Cor por Total de Casos de Leptospirose	30
Figura 9 – Disposição Geográfica do Grupo 1 com Graduação de Cor por Total de Casos de Leptospirose	30
Figura 10 – Disposição Geográfica do Grupo 1 com Graduação de Cor por Total de Casos de Leptospirose	31
Figura 11 – Disposição Geográfica do Grupo 1 com Graduação de Cor por Densidade Populacional	32

Lista de tabelas

Tabela 1 – Estatísticas do Subgrupo de Características Gerais dos Municípios do Grupo 0	28
Tabela 2 – Estatísticas do Subgrupo de Características Gerais dos Municípios do Grupo 1	29

Listas de símbolos

C_i	i-ésimo grupo.
k	Número de grupos.
x_j	j-ésimo objeto multidimensional.
n	número de objetos.
$d(x_i, x_j)$	Distância entre os objetos x_i e x_j .

Sumário

1	Introdução	9
1.1	Objetivo Geral	10
1.2	Objetivos Específicos	10
2	Referencial Teórico	11
2.1	Epidemiologia	11
2.2	Mineração de Dados	12
2.2.1	Coleta de Dados	12
2.2.2	Pré-processamento	13
2.2.3	Clusterização	13
2.2.3.1	Clusterização Hierárquica Divisiva	13
2.2.3.2	Clusterização Hierárquica Aglomerativa	13
2.2.3.2.1	Algoritmo <i>Single-Linkage</i>	14
2.2.3.2.2	Algoritmo <i>Complete-Linkage</i>	14
2.2.3.2.3	Algoritmo <i>Group Avarage-Linkage</i>	15
2.2.3.2.4	Algoritmo <i>Ward</i>	15
2.2.3.3	Clusterização Particional	16
2.2.3.3.1	K-Médias	16
2.2.4	Validação	17
2.2.5	Interpretação	18
2.2.5.1	Banco de Dados Geográficos	18
2.2.5.2	Armazenamento e Representação Vetorial	18
2.2.5.3	Armazenamento e Representação Matricial	19
3	Trabalhos Relacionados	20
3.1	Indicadores Socioeconômicos E De Saúde Da Atenção Básica Nos Municípios Da Região Metropolitana De Belo Horizonte	20
3.2	Perfil De Unidades Básicas De Saúde Quanto Às Ações De Rastreamento Do Câncer Do Colo Do Útero No Rio De Janeiro	21
3.3	Aspectos Epidemiológicos E Socioeconômicos Relacionados Aos Casos De Óbito Por Tuberculose No Estado De Mato Grosso Do Sul	21
3.4	Mineração de Dados Aplicada à Tuberculose nos Municípios do Estado de São Paulo	22
4	Metodologia	23
4.1	Mineração de Dados	23
4.1.1	Coleta dos Dados	23
4.1.2	Pré-processamento dos dados	24
4.1.3	Agrupamento dos dados	24

4.1.4	Validação dos resultados	25
4.1.5	Pós-processamento e Interpretação	25
4.2	Georreferenciamento e Visualizações	26
5	Resultados	27
5.1	Análises Estatísticas dos Agrupamentos	27
5.2	Visualizações dos Agrupamentos em Mapas	29
6	Discussão	33
6.1	Trabalhos Futuros	34
	Referências	35

1 Introdução

O estudo epidemiológico é importante para avaliar a ocorrência de doenças e como afetam a sociedade(BRITO, 2012). A Leptospirose é uma doença endêmica no Brasil, assim como na América Latina como um todo, com impacto na saúde pública (GENOVEZ, 2009). Tal doença é causada pela bactéria do gênero *Leptospira*, uma bactéria adaptável e ativamente móvel que objetiva animais domésticos e silvestres que, no entanto, acomete humanos como hospedeiro acidental (GENOVEZ, 2009). O contágio é atrelado a fatores ambientais, logo explicáveis por meio de condições das moradias, ocupações e lazer (FIGUEIREDO et al., 2001). Sendo assim, para essa rastreabilidade existem dados de transparência fornecidos por plataformas digitais, como o DATASUS¹ (SILVA, 2019), os quais foram utilizados na etapa de coleta e formação da amostra, que carregam dados nacionais sobre casos de doenças e até mesmo sobre ataques de animais peçonhentos. Como complemento à disponibilidade do DATASUS, o IBGE foi empregado na incorporação de indicadores socioeconômicos que dizem respeito à condição geral do município.

Partindo do conhecimento de que humanos não são qualificados para analisar grande quantidade de dados de forma manual, principalmente quando se trata de dados multivalorados (que carregam diversas informações sobre o paciente) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), a ciência da computação contribui com métodos de análise voltados para diversos contextos de pesquisa, incluindo a Mineração de Dados (JAIN; DUBES, 1988; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A análise por agrupamento é um dos métodos mais utilizados na Mineração de Dados (JAIN; DUBES, 1988) e foi escolhida como abordagem deste projeto de pesquisa com a finalidade exploratória de detecção de perfis municipais segundo os indicadores selecionados. Além disso, a base de dados formada para a pesquisa consiste em 291 municípios de São Paulo (cada um identificado pelo código do IBGE e representando uma linha única) de forma que são, individualmente, caracterizados por: número de casos por escolaridade, número de casos por nível de urbanização, número de casos por ocupação, densidade populacional do município e IDH-M, que definem as colunas da base. Para uma abordagem metódica, o processo *Knowledge Discovery in Databases* (KDD, do português: Descoberta de Conhecimento em Base de Dados), descrito no capítulo 4, oferece um padrão de Mineração de Dados envolvendo tarefas desde a coleta de dados até a obtenção de conhecimento (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Como complemento ao KDD foi escolhida a técnica de visualização dos agrupamentos por meio de mapas. Vale destacar que, entre os benefícios de se utilizar um Sistema de Informações Geográficas (SIGs) está a possibilidade de se observar grande volume de

¹ DATASUS:<http://siab.datasus.gov.br/DATASUS/index.php?area=0203id=29878153>

dados rapidamente, via unidades gráficas elementares (linhas, pontos, superfícies e volumes), no formato de paisagem digital ([LUCIA, 2017](#)). Com isso foi possível enriquecer os resultados da etapa de Mineração por meio da visualização dos grupos nas respectivas disposições geográficas e identificação de foco regional que independe do agrupamento efetuado. Nessa etapa a amostra foi importada pelo Sistema de Gerenciamento de Banco de Dados PostgreSQL e manipulada para formação de tabelas que geraram as visualizações presentes no capítulo: 5 para criação de visualizações utilizando o QGIS, um Sistema de Informação Geográfica de código aberto.

1.1 Objetivo Geral

Analizar dados digitais públicos para descrição epidemiológica através da mineração de dados, filtrar os dados em Banco de Dados Geográfico e criar visualizações (mapas).

1.2 Objetivos Específicos

- Coletar os dados de Leptospirose na série temporal disponível pelo DATASUS e complementar com características municipais disponíveis pelo IBGE;
- Unir as diferentes bases coletadas em um único DataFrame utilizando python para aplicação de métodos da biblioteca pandas;
- Pré-processar os dados por meio da eliminação de dados faltantes e normalizar os elemnts restantes;
- Aplicar K-Médias e os algoritmos de Clusterização Hierárquica;
- Avaliar todos resultados afim de selecionar: o melhor algoritmo sobre a condição de k grupoos e, por consequênciia, o melhor agrupamento;
- Efetuar pós-processamento por meio de medidas estatísticas para o próximo passo de interpretação;
- Interpretar as tabelas e gráficos resultantes;
- Por meio da interpretação produzir visualizações que destacam características dos municípios afim de esclarecer o contexto social e geográfico da amostra;
- Interpretar os mapas.

2 Referencial Teórico

Neste capítulo estão apresentados os conceitos necessários para compreensão desta pesquisa como epidemiológica e do contexto da MD. Encontra-se um foco nas técnicas da mineração que foram utilizadas, como o framework de processos e os diferentes algoritmos de clusterização. Por fim é apresentada a abordagem que envolve a visualização dos dados georreferenciados e que oferece apoio na etapa de extração de conhecimento da MD.

2.1 Epidemiologia

Há mais de 2000 anos Hipócrates observou que fatores ambientais influenciam na ocorrência de doenças (BRITO, 2012). Considerado esse um dos primeiros registros informais da epidemiologia. Mas foi apenas no final do século XIX, principalmente devido aos estudos de John Snow sobre a cólera, é que a prática de se comparar coeficientes de doenças em subgrupos populacionais se tornou comum (BRITO, 2012; FINE et al., 2013). Portanto, a epidemiologia que conhecemos atualmente é uma disciplina que usa métodos quantitativos para estudar a ocorrência de doenças nos humanos com a finalidade de definir estratégias de gestão que envolvem prevenção e controle de reincidência, se mostrando uma ciência fundamental para saúde pública (BRITO, 2012). Muitos problemas de saúde podem ser prevenidos por meio da mudança de hábitos, como a higienização das mãos ou a prática de “sexo seguro” (FINE et al., 2013). Essas conclusões partem do estudo sobre a proliferação de uma doença e, posteriormente, o levantamento de hipóteses sobre como preveni-la. Durante o estudo de John Snow algumas teorias se formaram quando foi feita a análise de dispersão geográfica da doença em relação a indicadores de saneamento (FINE et al., 2013), destacando a importância desse âmbito na pesquisa epidemiológica.

Pelo fato do crescente volume de dados sendo gerado, a Mineração de Dados foi proposta na década de 80 e ainda está em desenvolvimento para facilitar a busca pelo conhecimento (SOUZA; ZAIA, 2015). A busca em comum da epidemiologia e da Mineração de Dados por métodos regrados que orientam a pesquisa (BRITO, 2012; SOUZA; ZAIA, 2015) se desenvolveram ao ponto dos dois conceitos serem unidos no framework *KDD* como um método computacional nesta pesquisa em questão. Visando complementar a análise dos dados por meio de indicadores, o georreferenciamento dos dados é introduzido como etapa incremental para criação de visualizações do estudo em mapas.

2.2 Mineração de Dados

A mineração de dados é uma tecnologia generalizável que pode ser aplicada em qualquer tipo de dado, desde que tenha significado para o objetivo da aplicação (HAND, 2008). Independente do contexto aplicado, a ciência de dados visa manusear e transformar uma grande quantidade de dados brutos em resultados úteis para auxiliar em decisões (SANTOS, 2020), em que a organização dos dados em grupos de semelhança é uma das práticas mais comuns. Esta prática descritiva e exploratória, sem a rotulação prévia de classes, é chamada de Clusterização, o estudo formal de algoritmos e métodos para agrupamento e classificação de objetos (JAIN; DUBES, 1988). O resultado é uma organização simples em que as classes encontradas (*clusters*) demonstram coerência interna (padrão do agrupamento).

Esse contexto de multidisciplinaridade da ciência de dados exigiu um *framework* para padronizar a busca por informações relevantes (FAYYAD USAMA M E PIATETSKY-SHAPIRO, 1996). Por meio dessa motivação surgiu o processo de descoberta de conhecimento: *Knowledge Discovery in Databases* (*KDD*), um conjunto de passos para guiar a mineração de dados, como demonstrado na Figura 1:

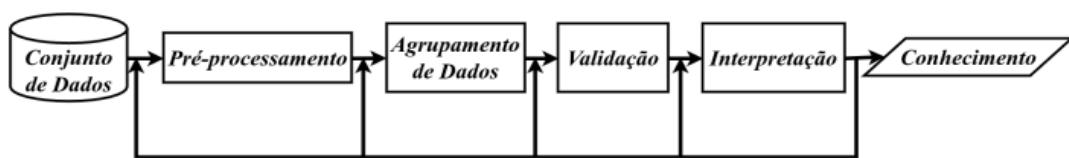


Figura 1 – Processo descrito pelo *KDD*. Foco em Agrupamento de Dados

(XU; WUNSCH, 2005 apud SANTOS, 2018)

Nota-se no *KDD* que os dados a serem tratados pelos algoritmos de agrupamento necessitam de pré-processamento, um tratamento para adequar os elementos para análise. Sendo assim os dados podem passar pela clusterização e logo após o modelo gerado é validado. Após a validação os resultados são interpretados para decisão de parar o ciclo da mineração ou então seguir para descoberta de conhecimento. Nas subsessões a seguir as etapas citadas podem ser conferidas com maior detalhamento teórico.

2.2.1 Coleta de Dados

Primeira etapa no processo de descoberta de conhecimento também conhecida como Seleção de Dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Nesse momento da pesquisa são definidas quais instâncias e elementos serão capturados e com quais características a serem analisadas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

2.2.2 Pré-processamento

Esta etapa pode ser definida como uma ou mais operações básicas para tratar dados faltantes (SANTOS, 2018; FAYYAD USAMA M E PIATETSKY-SHAPIRO, 1996), incoerentes, muitas vezes devido erros de inserção dos dados na base de origem (SANTOS, 2018) e até aproximar os intervalos de atributos distintos para que algoritmos não sejam enviesados.

Observa-se que o *KDD* é *framework* que guia o experimento, ou seja, o pré-processamento possivelmente não será efetuado apenas uma vez, diferentes técnicas costumam ser aplicadas para que os dados brutos se adéquem ao contexto (SANTOS, 2018; FAYYAD USAMA M E PIATETSKY-SHAPIRO, 1996).

2.2.3 Clusterização

A Clusterização costuma ser dividida em duas subcategorias: as particionais e as hierárquicas (JAIN; DUBES, 1988). A técnica de agrupamento hierárquico efetua junções, ou separações (abordagem aglomerativa e divisiva, respectivamente), em diferentes números de grupos ao decorrer do tempo de execução (XU; WUNSCH, 2005). Enquanto isso as abordagens particionais apresentam o resultado referente apenas ao número de agrupamentos pré-determinados (JAIN; DUBES, 1988). A seguir estão detalhadas as abordagens e os algoritmos utilizados no agrupamento da amostra em questão.

2.2.3.1 Clusterização Hierárquica Divisiva

Os algoritmos de clusterização hierárquica no geral possuem a vantagem da representação gráfica do processo de agrupamento que é chamada de Dendrograma (JAIN; DUBES, 1988), como pode ser visto na Figura 2:

Por sua vez, a abordagem divisiva dos algoritmos hierárquicos não se mostram tão eficientes quanto às aglomerativas. O caminho divisivo parte da amostra como uma única classe e efetua subdivisões, par a par, até que a amostra esteja totalmente unificada (JAIN; DUBES, 1988). Sendo assim, para um grupo de N objetos existem $2^{(N-1)}-1$ possibilidades de divisão. O que agrega alto custo computacional (EVERITT; LANDAU; LEESE, 2001).

2.2.3.2 Clusterização Hierárquica Aglomerativa

Os algoritmos de clusterização hierárquica aglomerativa, abordagem adotada para a pesquisa em questão, inicializam o número de grupos com o valor do número de objetos e possuem a condição de parada como $K = 1$, ao contrário dos algoritmos divisivos (XU; WUNSCH, 2005). Alguns dos algoritmos mais comuns com essa característica aglomerativa são: *Single-Linkage* (FLOREK et al., 1951 apud SANTOS, 2018), *Complete-Linkage* (JAIN; DUBES, 1988), *Group Avarage-Linkage* (AGGARWAL CHARU C E REDDY,

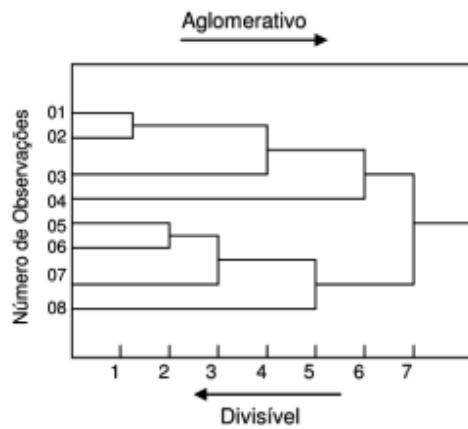


Figura 2 – Dendrograma do algoritmo hierárquico. Sentidos Aglomerativo e Divisível
(NASCIMENTO, 2011 apud SANTOS, 2020), Adaptada

2013 apud SANTOS, 2018) e *Ward* (método de variância mínima)(AGGARWAL CHARU C E REDDY, 2013 apud SANTOS, 2018), em que apenas os três últimos algoritmos citados foram utilizados nessa pesquisa.

2.2.3.2.1 Algoritmo *Single-Linkage*

A aglomeração dos objetos nesse algoritmo ocorre de acordo com a distância mínima entre eles, se apresentando como um dos critérios mais simples de junção (FLOREK et al., 1951 apud SANTOS, 2018). Considerando o princípio aglomerativo (ausência de grupos vazios) e de que os conjuntos formados são sempre disjuntos, não possuem objetos em comum, temos (Equação 2.1(FLOREK et al., 1951 apud SANTOS, 2018)):

$$d(C, C') = \min_{x \in C, y \in C'} d(x, y) \quad (2.1)$$

Em que, $d(C, C')$ retorna a distância entre dois grupos cuja foi calculada a partir dos objetos mais próximos entre eles. Na Equação 2.1, x é um objeto do grupo C e y do grupo C' . Sendo assim, esse algoritmo se apresentou sensível a ruídos em suas aplicações, mas é eficiente na identificação de agrupamentos com disposições gráficas arbitrárias (JAIN; DUBES; JOHNSON, 1988, 1967 apud SANTOS, 2018), não globulares/não esféricos, por exemplo, mas foi isento do processo de clusterização.

2.2.3.2.2 Algoritmo *Complete-Linkage*

O algoritmo *Complete-Linkage* se aproxima do *Single-Linkage* na forma de medir a distância entre os grupos (Equação 2.2 (JAIN; DUBES, 1988)):

$$d(C, C') = \max_{x \in C, y \in C'} d(x, y) \quad (2.2)$$

Nota-se que a distância é medida de acordo com a medida de dissimilaridade entre os objetos mais distantes de cada grupo já formado. Os resultados deste algoritmo demonstraram uma vantagem sobre o *Single-Linkage* na identificação de grupos com formato globular e se mostrou mais resistente a ruídos ([AGGARWAL CHARU C E REDDY; JAIN; DUBES, 2013, 1988 apud SANTOS, 2018](#)).

2.2.3.2.3 Algoritmo *Group Avarage-Linkage*

O *Group Avarage-Linkage* calcula a distância entre os grupos de acordo com a média de pares de objetos entre grupos (Equação 2.3 ([JAIN; DUBES, 1988](#))):

$$d(C, C') = \frac{1}{|C||C'|} \sum_{x \in C, y \in C'} d(x, y) \quad (2.3)$$

Observa-se que a combinação dos pares possíveis de objetos de grupos distintos é calculada pela multiplicação da quantidade de elementos no grupo C com a quantidade de elementos no grupo C'. Esse algoritmo tem uma limitação a contextos em que os grupos são globulares, mas apresenta eficiência na descrição de amostras com ruídos e *outliers* ([JAIN; DUBES, 1988 apud SANTOS, 2018](#)).

2.2.3.2.4 Algoritmo *Ward*

O algoritmo *Ward*, também chamado de Método da Variância Mínima, calcula a soma de quadrados (Equação 2.4 ([JOE, 1963](#))), como medida intragrupo ([HAIR et al., 2009](#)).

$$J = \sum_{i=1}^k \sum_{x_j \in C_i} d(x_j, \bar{x}_i)^2 \quad (2.4)$$

Sendo $d(\cdot)$ a Distância Euclidiana ¹, (\bar{x}_i) o centroide do grupo sendo analisado e k o número total de grupos ([JOE, 1963](#)), temos que uniões de teste são feitas para comparação dos valores anteriores e posteriores a operação. Dessa maneira, o grupo resultante com o menor aumento da soma quadrada terá sua união oficializada ([HAIR et al., 2009](#)).

¹ Nota-se que para o uso da Distância Euclidiana é necessário extrair a raiz quadrada do resultado do fato que pode reduzir o desempenho do algoritmo([HAIR et al., 2009](#))

2.2.3.3 Clusterização Particional

O conceito mais aplicado é o de *hard partitional clustering*, em que cada elemento da base pertence a somente um grupo (XU; WUNSCH, 2005). O modelo *hard* é utilizado neste projeto em questão e possui os seguintes princípios (XU; WUNSCH, 2005):

1. $C_i \neq \phi, i = 1, \dots, K;$
2. $\bigcup_{i=1}^K C_i = X;$
3. $C_i \cap C_j = \phi, i, j = 1, \dots, K$ em que $i \neq j.$

Sendo X a amostra de entrada, e C um agrupamento. E dentre os algoritmos mais comuns que seguem essa restrição geral temos o K-Médias (*K-Means*) (MACQUEEN; JAIN, 1967, 2010 apud SANTOS, 2018), apresentado a seguir.

2.2.3.3.1 K-Médias

Para fins de comparação de resultados foi selecionada uma técnica de agrupamento particional, o algoritmo K-Médias(*K-Means*). A diferença entre os métodos hierárquicos e os particionais consiste no fato dos particionais terem como hiperparâmetro um número K de grupos pré-definidos que limitará o retorno a essa divisão (ŘEZANKOVÁ HANA E EVERITT, 2009). A quantidade K de grupos define o número de pontos a serem distribuídos como centros dos grupos. As coordenadas dos pontos centrais podem ser selecionadas de acordo com diversos métodos, como a seleção de K pontos aleatórios da amostra ou seleção dos K primeiros pontos da amostra (ŘEZANKOVÁ HANA E EVERITT, 2009). O algoritmo visa encontrar um agrupamento cujo Erro Quadrático entre a média do grupo e seus pontos internos tenha sido minimizado (JAIN, 2010). Tendo $X = x_i$, em que i varia no número de objetos da amostra ($i = 1, \dots, n$) e K sendo o número pré-definido de agrupamentos, $C = c_k, k = 1, \dots, K$, temos a seguinte análise (Equação 2.5):

$$J(C_k) = \sum_{x_i \in C_k} \|x_i - \mu_k\|^2. \quad (2.5)$$

Observa-se que cada elemento é analisado para inserção em um grupo C_k . Além disso é necessário que esses elementos também sejam testados perante os outros grupos em formação ($C = c_k, k = 1, \dots, K$), logo temos a Equação 2.6:

$$J(C) = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2. \quad (2.6)$$

Por isso o K-Médias é considerado um algoritmo guloso, em que seu desempenho depende fortemente do contexto e da inicialização dos centroides (JAIN, 2010). Os passos principais do algoritmo podem ser descritos como (JAIN; DUBES, 1988; JAIN, 2010):

1. Selecionar e efetuar a inicialização dos K centroides;
2. Inserir um objeto da amostra e atribuí-lo ao centroide mais próximo;
3. Ajustar o centroide por meio da média das distâncias dos objetos internos ao grupo;
4. Repetir 2 e 3 até que os grupos estejam estáveis.

A métrica de distância mais utilizada é a Euclidiana. Como consequência o formato dos agrupamentos do algoritmo em que se utiliza essa métrica é de corpos globulares/circulares (JAIN, 2010).

2.2.4 Validação

A validação é um meio de comparar métodos aplicados sobre diferentes pré-processamentos ou hiperparâmetros distintos e também para algoritmos diferentes (JAIN; DUBES; XU; WUNSCH, 1988, 2005 apud SANTOS, 2018) Para cada aplicação da validação é necessário considerar a natureza dos dados, portanto as medidas de dissimilaridade utilizadas, e a classe do algoritmo (SANTOS, 2018; ROUSSEEUW, 1987). Considerando que a pesquisa utiliza algoritmos descritivos hierárquicos e particional, sobre a métrica Euclidiana para distâncias, o método *Silhouette Score* (Pontuação da Silhueta) apresenta-se eficiente (ROUSSEEUW, 1987). Para que essa validação seja feita o método necessita dos dados rotulados por um método de agrupamento e da matriz de distâncias entre os objetos rotulados (ROUSSEEUW, 1987). Considerando um grupo inicial A e um elemento arbitrário i, tem-se:

- 1: Cálculo da dissimilaridade média de i para outros elementos intragrupo ($a(i)$);
- 2: Cálculo da dissimilaridade média de i para elementos dos grupos externos;

Supõe-se um grupo C, em que tem-se: $d(i,C)$, sendo C qualquer agrupamento diferente de A. Por fim, temos:

- 3: $b(i) = \min(d(i, C))$, com A sendo qualquer grupo diferente de C.

O terceiro passo diz que para $b(i)$ será retornada a menor dissimilaridade média de i para os grupos externos. Por fim, o cálculo que resulta um valor objetivo para validação é (Equação 2.7):

$$s(i) = \frac{b(i) - a(i)}{\text{Max}\{a(i), b(i)\}} \quad (2.7)$$

A silhueta $s(i)$ está no intervalo numérico $[-1,1]$, em que $s(i) = 0$ quando o número de agrupamentos é igual a um. Quando uma silhueta próxima a um é retornada significa que o elemento de um grupo A está mais próximo dos elementos de seu próprio grupo ($a(i)$) do que está de outro grupo próximo a A ($b(i)$). Ao contrário, teríamos um valor próximo a -1.

2.2.5 Interpretação

Este é o último passo do KDD em que os resultados são documentados e ocorre a decisão de retorno a um dos passos anteriores para refinamento dos métodos utilizados, quando é o caso de resultados insatisfatórios ([FAYYAD USAMA M E PIATETSKY-SHAPIRO, 1996](#)). Em que a visualização dos retornos pode ser praticada de diversas formas, como nesse trabalho em que além do registro em tabelas e gráficos os objetos de pesquisa já classificados foram expostos em contexto geográfico. ([FAYYAD USAMA M E PIATETSKY-SHAPIRO, 1996](#)).

2.2.5.1 Banco de Dados Geográficos

Um Sistema Gerenciador de Banco de Dados (SGBD) possui a capacidade de servir a diversas aplicações pois encapsula seus bancos de acordo com as estruturas desejadas e suas restrições ([FONSECA, 2020](#)). Neste trabalho de conclusão de curso foi utilizado o SGBD PostgreSQL com a extensão para dados geográficos disponível (PostGIS), em que se torna possível, além do gerenciamento e recuperação dos dados, a manipulação de informações geográficas. Sendo assim, o banco de dados passa a trabalhar não só com dados alfanuméricos mas também com dados espaciais que podem ser representados nas formas matriciais ou vetoriais ([FITZ, 2018 apud OLIVEIRA, 2021](#)), assim como expostos na Figura 3:

2.2.5.2 Armazenamento e Representação Vetorial

A estrutura vetorial é uma composição de pontos ou conjunto de pontos interligados que utilizam coordenadas geográficas para serem representados ([FITZ, 2018 apud OLIVEIRA, 2021](#)). Sendo assim, é possível representar contornos de municípios e estados, como utilizado nesse trabalho em questão, pontos de interesse (como centroides de outras geometrias), viadutos por meio de linhas, etc. ([FITZ, 2018 apud OLIVEIRA, 2021](#))

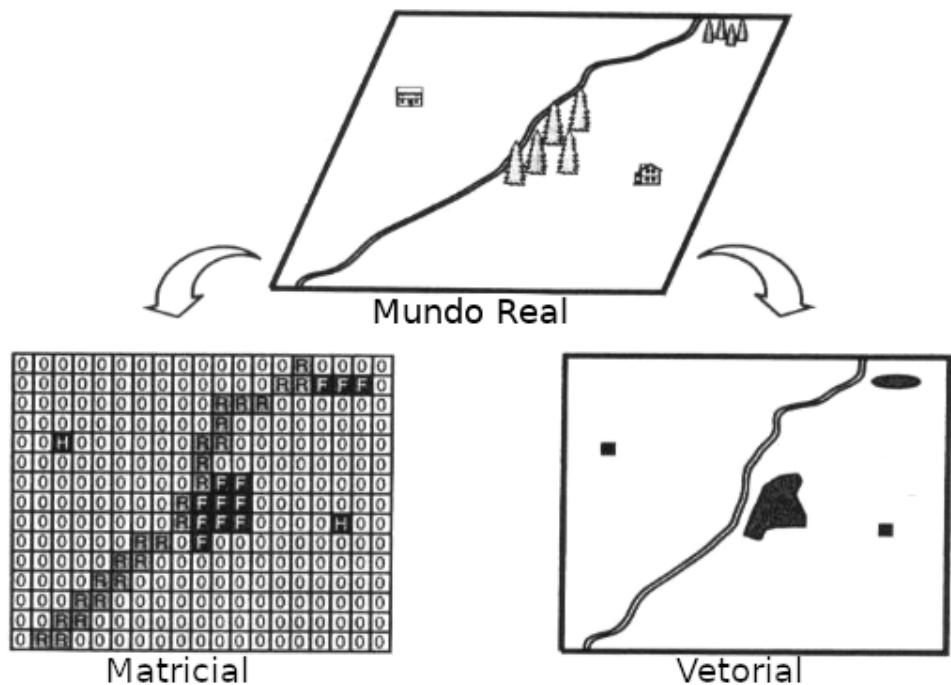


Figura 3 – Comparaçāo Entre Representaçāo Matricial e a Vetorial

(NASCIMENTO, 2011 apud SANTOS, 2020), Adaptada

2.2.5.3 Armazenamento e Representação Matricial

Essa estrutura é uma matriz de pixels, também conhecida como *raster*, em que cada encontro de linha com coluna é formado um par de coordenada (FITZ, 2018 apud OLIVEIRA, 2021). Sendo assim a relação com dados alfanuméricos, utilizados para descrição dos dados espaciais (OLIVEIRA, 2021), se torna dificultada devido a atribuição ser *pixel a pixel*.

Para visualização dos dados foi utilizado o Sistema de Informações Geográficas (SIG) QGIS, um software de código aberto que possibilita a conexão com o banco de dados e manipulações em SQL das bases conectadas para criação de visualizações. Em geral o SIG é uma definição de sistema de informação que pode armazenar atributos descritivos e geometrias de diversos dados geográficos distintos, que permite manipulações para geração de camadas e criação de visualizações (CASANOVA, 2005 apud OLIVEIRA, 2021).

3 Trabalhos Relacionados

Abaixo estão artigos relacionados com o objetivo deste trabalho de conclusão de curso. Serão citados os títulos referentes e em seguida uma descrição deles e sobre como se relacionam com este Trabalho de Conclusão de Curso (TCC).

3.1 Indicadores Socioeconômicos E De Saúde Da Atenção Básica Nos Municípios Da Região Metropolitana De Belo Horizonte

Essa pesquisa ([CAMPOS et al., 2012](#)) teve como objetivo agrupar municípios do estado de Minas Gerais por indicadores socioeconômicos e então relacioná-los por indicadores de saúde com a finalidade de apoiar tomada de decisões do governo perante áreas sensíveis. A amostra consistiu em 33 municípios ao redor da capital (Belo Horizonte) e foi recolhida a partir dos Sistemas de Informações em Saúde (SIS) do Departamento de Informática do Sistema Único de Saúde – DATASUS. Foram utilizados dois índices gerais de análise socioeconômica: IDH e Índice de Gini; e um indicar específico para o caso do estado de MG: Índice Mineiro de Responsabilidade Social. Como indicadores de saúde utilizou-se: indicadores de saúde geral, que indicam cuidados médicos em diferentes âmbitos por meio das variáveis independentes, e de saúde bucal. A segmentação dos municípios foi feita utilizando o método de Clusterização Hierárquica, pelas variáveis analisadas, demonstrou que a capital mineira possui um cenário muito destoante do resto da amostra. Além disso foi gerada base para análise estatística por meio do t-Student, que revelou que locais menos favorecidos(grupo com IDH baixo) recebem mais atenção do governo por meio da Estratégia Saúde da Família. Apontam também que o grupo que possui maior IDH global possui municípios com alto Índice de Gini(mede desigualdade de renda) e que não recebem atenção governamental para saúde. Os autores indicaram que a amostra oferece limitação da análise, mas que foi uma escolha adequada para possibilitar o uso do método de segmentação hierárquico. A amostra utilizada, que representa aproximadamente 3,869% do número total de municípios de MG (853 municípios) se mostra limitada, portanto, a aplicação do trabalho de conclusão de curso em questão tem a intenção de exercitar o algoritmo hierárquico com uma amostra consideravelmente maior. Por outro lado, o trabalho citado contribui com a sugestão de indicadores socioeconómicos e exemplificação do comportamento do algoritmo de Clusterização Hierárquica nesse contexto de epidemiologia.

3.2 Perfil De Unidades Básicas De Saúde Quanto Às Ações De Rastreamento Do Câncer Do Colo Do Útero No Rio De Janeiro

Este trabalho ([AHMED et al., 2014](#)) tem como objetivo traçar o perfil das unidades básicas de saúde de três regiões do estado do Rio de Janeiro. Para isso foi reunida uma amostra de 535 dessas unidades por meio do Sistema de Informação do Câncer do Colo do Útero (SISCOLO). Para análise foi utilizado *K-means* em comparação com método hierárquico completo, utilizando-se percentuais das variáveis relacionadas à: organização do serviço, resultado do exame citopatológico e condições de saúde das mulheres usuárias do programa. Após testes de aplicação, por meio da linguagem R (versão 3.0.2), foi decidido que os métodos seriam: método hierárquico com agregação completa e o k-means, ambos utilizando a distância euclidiana. Os resultados obtidos foram próximos e, sendo assim, nenhum deles foi invalidado e observou-se que as condições socioeconômicas e organizacionais parecem ter influência no desempenho das ações de um programa de rastreamento. A pesquisa citada contribui para este trabalho principalmente como exemplo de aplicação de Clusterização Hierárquica que utiliza uma amostra considerável de dados reais (535 unidades analisadas) em relação a outras pesquisas, além do fato de ocorrer a comparação entre métodos de segmentação (método hierárquico e *Kmeans*).

3.3 Aspectos Epidemiológicos E Socioeconômicos Relacionados Aos Casos De Óbito Por Tuberculose No Estado De Mato Grosso Do Sul

A amostra do estudo de ([BALDAN; NUNES; ANDRADE, 2018](#)) consiste na união de dados do Sistema de Infomração de Mortalidade (SIM), Sistema de Infomração de Agravos de Notificação (SINAN) e Instituto Brasileiro de Geografia e Estatística (IBGE) para olhar multisectorial que envolve aspectos econômicos, sociais e ambientais. A partir da aplicação do método *Kmeans*, foram retornados 5 grupos de municípios em que em cada grupo existiam municípios com convergência do Coeficiente de Incidência, Coeficiente de Mortalidade, IDH e Índice de Gini. Os grupos de municípios foram identificados em mapa, por meio do *Terra View 4.2.2*, para melhor visualização e interpretação do agrupamento. Apesar dos autores terem utilizado softwares para aplicação do *Kmeans* e para formação do mapa, existe contribuição por parte da lógica utilizada na união das bases de dados, rumo a formação de cenários socioeconômicos, e no tratamento e normalização dos dados para a técnica de clusterização no contexto da saúde.

3.4 Mineração de Dados Aplicada à Tuberculose nos Municípios do Estado de São Paulo

A motivação da pesquisa de (SANTOS, 2020) é a alta incidência da Tuberculose em alguns municípios do Brasil, fato que se mostra conflitante pela doença ser tratável e poder ser prevenida por meio da vacinação. Devido a Tuberculose ser uma doença de contaminação direta e estar ligada a fatores socioeconômicos, o pesquisador aproveitou as bases de dados: SEADE, IBGE e DATASUS, com a finalidade de efetuar o estudo epidemiológico do estado de São Paulo por meio da computação. Os algoritmos escolhidos para esse estudo se diferem dos adotados para o trabalho de conclusão de curso em questão, porém o método de análise se assemelha devido a abordagem de mineração de dados para aproximação de incidência e fatores socioeconômicos lastreado no *KDD*, tendo como foco a análise descritiva, com a incorporação da técnica de mapeamento geográfico dos casos. O algoritmo de clusterização foi o *AGNES*, voltado para análise de séries temporais. A validação dos resultados foi efetuada com índice de *Silhouette* e coeficiente de correlação cophenético, um coeficiente facilitado pela aplicação da clusterização hierárquica na linguagem de programação R. A análise estatística espacial dos dados já classificados evidenciou municípios fora da lista de municípios de controle prioritário da doença. A pesquisa também contribui com mapas referentes aos agrupamentos gerados em que os municípios se apresentam em visualização por indicadores socioeconômicos utilizados na pesquisa por meio de cores discriminantes para grupo ou variação de índices.

4 Metodologia

Esta pesquisa em questão tem caráter descritivo e exploratório com foco na aplicação de abordagens computacionais para geração de informações a partir de dados públicos de saúde. Essa abordagem epidemiológica foi iniciada com a Mineração de Dados por meio dos algoritmos de agrupamento: Clusterização Hierárquica e K-Médias, utilizando *Python* na versão 3.7.15 e no ambiente de desenvolvimento *Google Colab*. Assim que foi possível obter os rótulos dos dados (classificação em dois grupos e nomeados como: *Grupo 0* e *Grupo 1*) a amostra ganhou caráter geográfico e foi mapeada nos limites do estado de São Paulo.

4.1 Mineração de Dados

O processo adotado foi a Descoberta de Conhecimento em Conjunto de Dados (*Knowledge Discovery in Databases – KDD*) em que, partindo de uma base de dados, nesse caso, estruturados (dados representados por meio de bancos de dados relacionais ou planilhas ([XU; WUNSCH, 2005](#)), e, como citado no capítulo 2, define-se basicamente cinco etapas: coleta de dados, pré-processamento, prática de técnica(s) de agrupamento de dados, validação e a interpretação dos dados. Nota-se por meio da Figura 1 que essas etapas podem ser executadas de forma cíclica como persistência para formação de conhecimento ([XU; WUNSCH, 2005 apud SANTOS, 2018](#)).

4.1.1 Coleta dos Dados

Na etapa de coleta de dados foram utilizadas duas fontes digitais governamentais de dados para formação da base a ser pesquisada. A partir do TABNET/DATASUS obteve-se a relação de casos por: escolaridade, ocupação durante contágio (durante lazer, em casa, trabalho ou outros) e região (urbana, periurbana, rural ou outras) em que cada município é representado como uma linha de uma tabela e as colunas são as características com os totais de casos de 2007-2020. Já para fins de contextualização dos casos com o município, o IBGE foi utilizado para acrescentar os indicadores de Densidade Populacional e IDH municipal (IDH-M), baseado no Produto Interno Bruto *per capita* a informações como: educação, taxa de escolaridade e à saúde. Além disso, foi coletada uma lista de códigos municipais para padronização dos 293 municípios analisados.

4.1.2 Pré-processamento dos dados

Com a conclusão da coleta de dados inicia-se o pré-processamento para neutralizar valores inconsistentes ou com intervalos numéricos discrepantes que podem afetar negativamente o processo de agrupamento (HAND, 2008 apud SANTOS, 2018). Nessa etapa os municípios que não tiveram o código correspondente na tabela do IBGE foram descartados, restando 291 elementos. Tendo padronizado os municípios então os dados foram normalizados segundo o método *Z-score*¹, uma vez que a contagem de casos possuía intervalos numéricos discrepantes, como no caso da densidade populacional do IDH-M. Essa técnica de normalização mantém a característica dos dados por ser baseada na média e desvio padrão, mas aproxima o intervalo numérico geral das colunas.

4.1.3 Agrupamento dos dados

A etapa de agrupamento consiste basicamente na extração de padrões dos dados tratados (CARVALHO; MILANI, 2013). Os algoritmos k-Médias e Clusterização Hierárquica foram escolhidos devido uma quantidade relevante de resultados positivos na literatura, o que leva a um suporte de consulta para o processo de pesquisa, e pelo fato de ambos serem métodos de segmentação não supervisionados. Para que os algoritmos segmentem a amostra é necessário atribuir um número K de grupos, uma medida de dissimilaridade (métrica de distância entre dois vetores multidimensionais) e, no caso do método hierárquico, algoritmos que vão definir como as métricas de distância serão interpretadas para o agrupamento.

Quando se trata do K-Médias (implementação do *scikit learn*) o K representa o número de centroides distribuídos com certa aleatoriedade e que se ajustam durante a execução por meio da média das distâncias dos objetos da amostra conforme são inseridos. Para inicialização do algoritmo foi selecionado o método *kmeans++*, que usa uma distribuição de centroides baseada na probabilidade e a medida de distância adotada foi a Distância Euclidiana. Assim como no método hierárquico.

Para a clusterização hierárquica (implementação do *scikit learn*), foram utilizados os métodos de segmentação: *Complete-Linkage*, *Average-Linkage* e *Ward*. Os algoritmos estão disponíveis nas formas divisiva e aglomerativa, mas a pesquisa aborda apenas a implementação aglomerativa por agregar menos complexidade computacional. Quando Aglomerativo, o método cria um grupo para cada elemento da amostra e inicia-se um processo de uniões consecutivas entre os grupos de acordo com a similaridade do objeto com os membros do grupo, como visto no capítulo 2, sendo que a condição de parada foi definida com valores pré-definidos de K.

Conforme o desvio padrão dos resultados ficou elevado, os mapas auxiliam a ana-

¹ <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.zscore.html>

lisar a dispersão dos indicadores e a representar as informações de maneira fidedigna.

4.1.4 Validação dos resultados

Os resultados desta etapa são rapidamente analisados de forma intuitiva e caso nenhuma inconsistência clara seja encontrada então uma métrica de validação é necessária. A métrica eleita foi o Coeficiente de Silhueta (*Silhouette Score*, implementada pela *scikit learn*) que verifica a coesão interna do cluster em relação aos objetos externos a ele. Sendo assim, a Silhueta foi aplicada para as seguintes combinações de aplicação: Clusterização Hierárquica e os métodos de agrupamentos utilizados, e K-Médias de acordo com valores de K (número de grupos) de 2 a 7, como pode ser visto no diagrama abaixo (Figura 4):

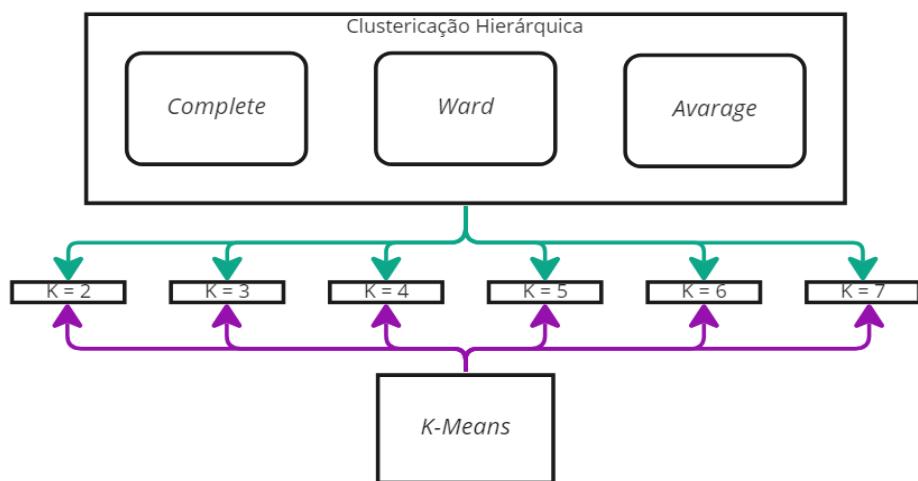


Figura 4 – Combinações dos Algoritmos com os Números de Grupos (Ks) para Validação.
Fonte: próprio autor

Como os algoritmos tiveram o mesmo valor de Silhueta para K=2 (dois grupos), em que a validação classificou como o melhor número para agrupamento dos dados, a abordagem estatística do *boxplot* (implementação da *matplotlib*) foi utilizada para quantificação de *outliers*, quantificação dos elementos superiores e inferiores aos quartis. Com isso o método Hierárquico aplicado com *Complete-Linkage*, que obteve o menor número de *outliers* e maior silhueta, foi eleito como o melhor retorno para o contexto da pesquisa.

4.1.5 Pós-processamento e Interpretação

O Pós-processamento é uma etapa para se verificar até que ponto a extração de características foi útil para o problema em questão (CARVALHO; MILANI, 2013). Por isso o resultado do algoritmo Complete-Linkage, com dois grupos, foi analisado por meio da correlação para seleção de municípios bem correlacionados no contexto geográfico (região urbana, periurbana ou rural), de ocupação durante a contaminação, escolaridade,

e contexto socioeconômico. Para cada um desses subgrupos de características foram selecionados municípios com 80% ou mais de correlação e para fins de contestação foram contabilizados municípios com 20% ou menos de correlação. Os representantes (objetos com 80% ou mais de correlação) foram analisados perante um método da biblioteca pandas chamado *describe*, que permitiu análise dos valores da média e desvio padrão dos valores de cada subgrupo de características.

O produto dessa análise estatística consiste na descrição do perfil de cada grupo. Essa descrição foi comparada com o *describe* aplicado ao *Grupo 0* e ao *Grupo 1* em suas totalidades de municípios para verificação e generalização da interpretação de cada *cluster*.

4.2 Georreferenciamento e Visualizações

Com a finalidade de mapear os dados e dar contexto geográfico a eles, como a visualização de vizinhança dos municípios de cada grupo, os dados já rotulados foram enriquecidos com suas respectivas características geográficas. A base de dados em extensão “.csv” foi importada pelo PostgreSQL, versão 14.6, e conectada ao QGIS, Versão 3.28.

PostgreSQL é um Sistema de Gerenciamento de Banco de Dados Relacional (SGBD) desenvolvido pela Universidade da Califórnia. Utilizando essa ferramenta os dados foram filtrados para separar o *Grupo 0* do *Grupo 1* e tiveram os limites municipais² atribuídos como tipo de dados “geometry”.

O QGIS³ por sua vez é um software SIG (Sistema de Informação Geográfica) profissional e de código aberto. Uma vez conectado à base de dados da pesquisa foi possível criar visualizações que diferenciassem ambos os grupos identificados e também permitiu a manipulação das geometrias para que as cores representativas variassem gradativamente de acordo com atributos como total de casos e valores de densidade populacional.

² <https://portaldemapas.ibge.gov.br/portal.phphomepage>

³ https://qgis.org/pt_BR/site/

5 Resultados

Esta seção é voltada para exposição dos resultados do processo de descoberta de conhecimento (*KDD*) e das visualizações geográficas dos dados.

5.1 Análises Estatísticas dos Agrupamentos

Pode-se observar na figura abaixo (5) o contraste entre os valores médios de casos por características do *Grupo 0* e do *Grupo 1*:

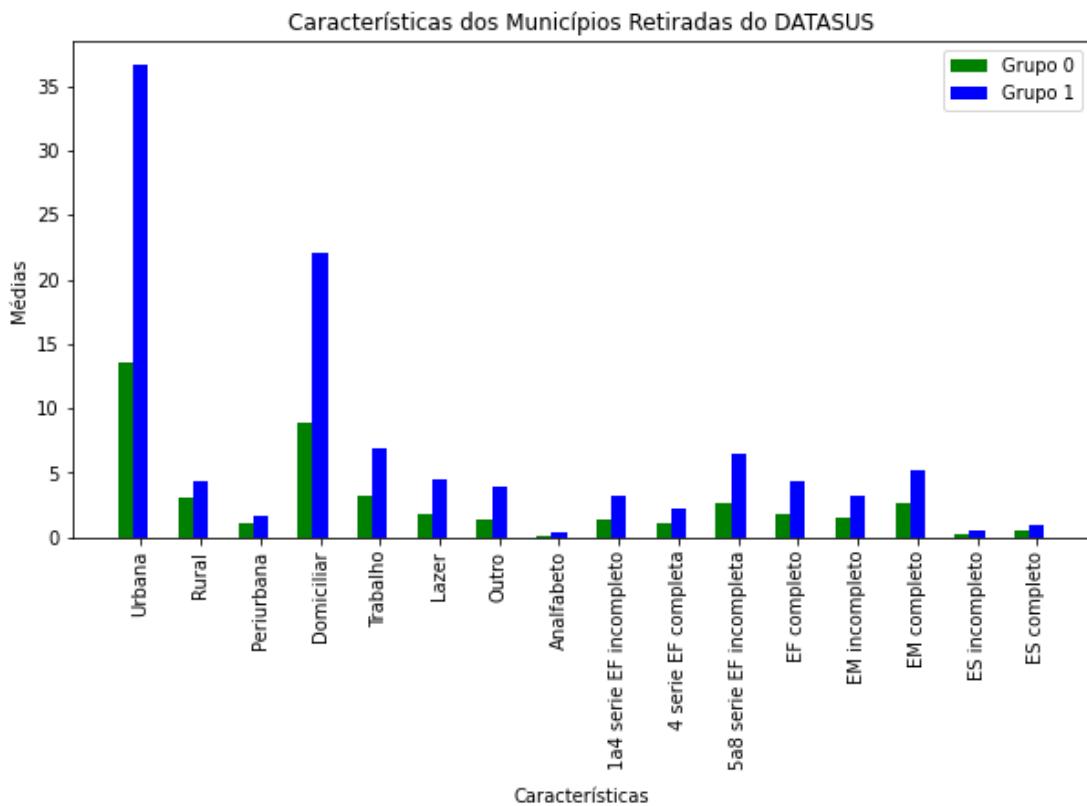


Figura 5 – Subgrupo de Características extraídas do DATASUS e seus valores médios para cada agrupamento.

Os dados apresentados na Figura 5 foram decisivos para se traçar o perfil dos agrupamentos encontrados e diferenciálos. Para que os valores das médias pudessem ser analisados com mais clareza a representação do desvio padrão foi adicionada em outro quadro (Figura 6). Abaixo podemos verificar as linhas de desvio padrão sobre as características, em que se tem noção da consistência dos grupos para tais dados:

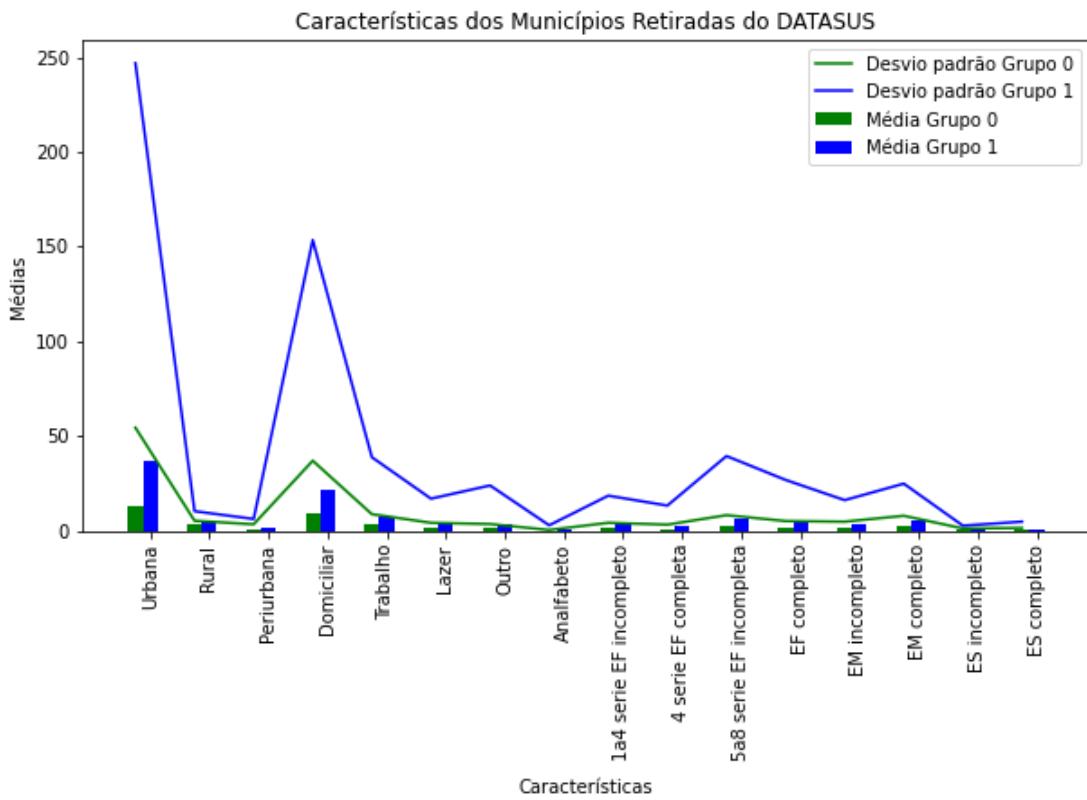


Figura 6 – Subgrupo de Características extraídas do DATASUS, seus valores médios e desvio padrão para cada agrupamento.

Observa-se também a seguinte análise estatística sobre as características municipais: Total de casos de Leptospirose, Densidade Populacional e IDH-M. A Tabela 1 é relativa ao *Grupo 0* que contém 187 municípios, enquanto a tabela 2 é relativa ao *Grupo 1*, com 104 municípios:

Grupo 0	Total de Casos	Densidade Populacional	IDH-M
Média	22.898396	625.072941	0.703676
Desvio Padrão	68.198834	1830.488965	0.037836
Valor Mínimo	1.000000	7.680000	0.561500
Valor Máximo	748.000000	13704.770000	0.783000

Tabela 1 – Estatísticas do Subgrupo de Características Gerais dos Municípios do Grupo 0

Grupo 1	Total de Casos	Densidade Populacional	IDH-M
Média	52.096154	731.229231	0.711375
Desvio Padrão	302.603028	1952.883312	0.043209
Valor Mínimo	1.000000	8.630000	0.609500
Valor Máximo	3075.000000	13176.770000	0.841000

Tabela 2 – Estatísticas do Subgrupo de Características Gerais dos Municípios do Grupo 1

De forma que foi feita a apresentação da análise estatística dos resultados pode-se partir para a disposição das visualizações dos resultados por meio de paisagens digitais com a finalidade de tornar palpável a dispersão das informações no mundo real.

5.2 Visualizações dos Agrupamentos em Mapas

A listagem dos municípios de cada agrupamento pode ser encontrada em repositório no *GitHub*¹, assim como o *backup* do banco de dados PostgreSQL em extensão ".csv". Abaixo, na Figura 7, tem-se o mapa de dispersão dos agrupamentos em que pode-se verificar a situação de vizinhança entre municípios da mesma classe ou não:

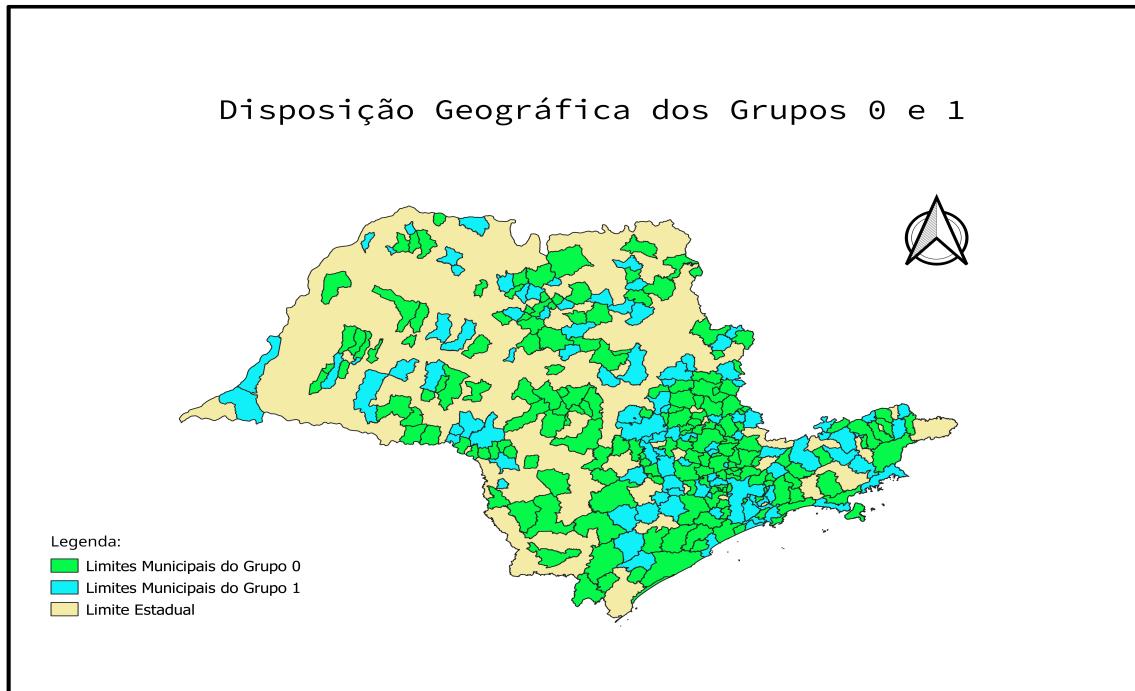


Figura 7 – Disposição Geográfica dos Grupos 0 e 1. Fonte: próprio autor

A seguir (nas Figuras 8 e 9) estão apresentadas as dispersões dos grupos submetidos às representações graduadas de cores para identificação da variação de casos por município:

¹ <https://github.com/patriani/EpidemiologiaLeptospiroseSP>

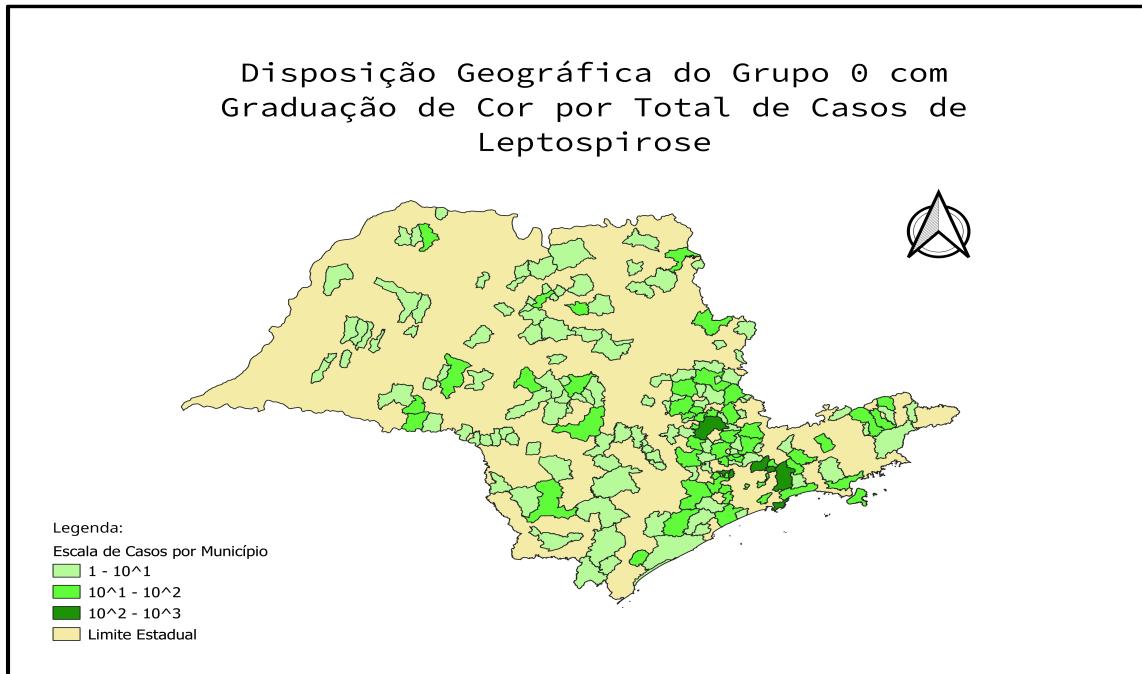


Figura 8 – Disposição Geográfica do Grupo 0 com Graduação de Cor por Total de Casos de Leptospirose. Fonte: próprio autor

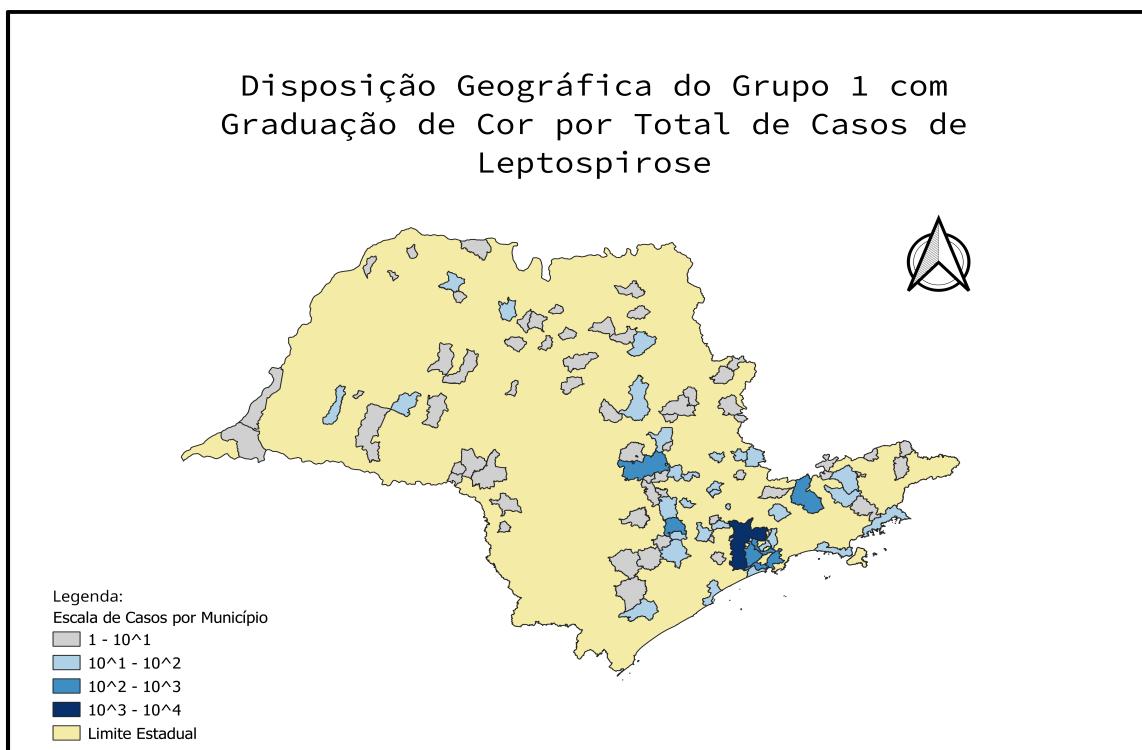


Figura 9 – Disposição Geográfica do Grupo 1 com Graduação de Cor por Total de Casos de Leptospirose. Fonte: próprio autor

Nas Figuras 8 e 9 foram apresentados os agrupamentos separados para noção de vizinhança municipal intragrupo em tentativa de evidenciação de padrões geográficos específicos de cada classe. E por fim, na Figura 10 e 11 são apresentados os mapas de dispersão dos *Grupos 0 e 1* com as cores representativas variando em: Total de casos e Densidade Populacional Municipal (*Habitantes/Km²*), respectivamente:

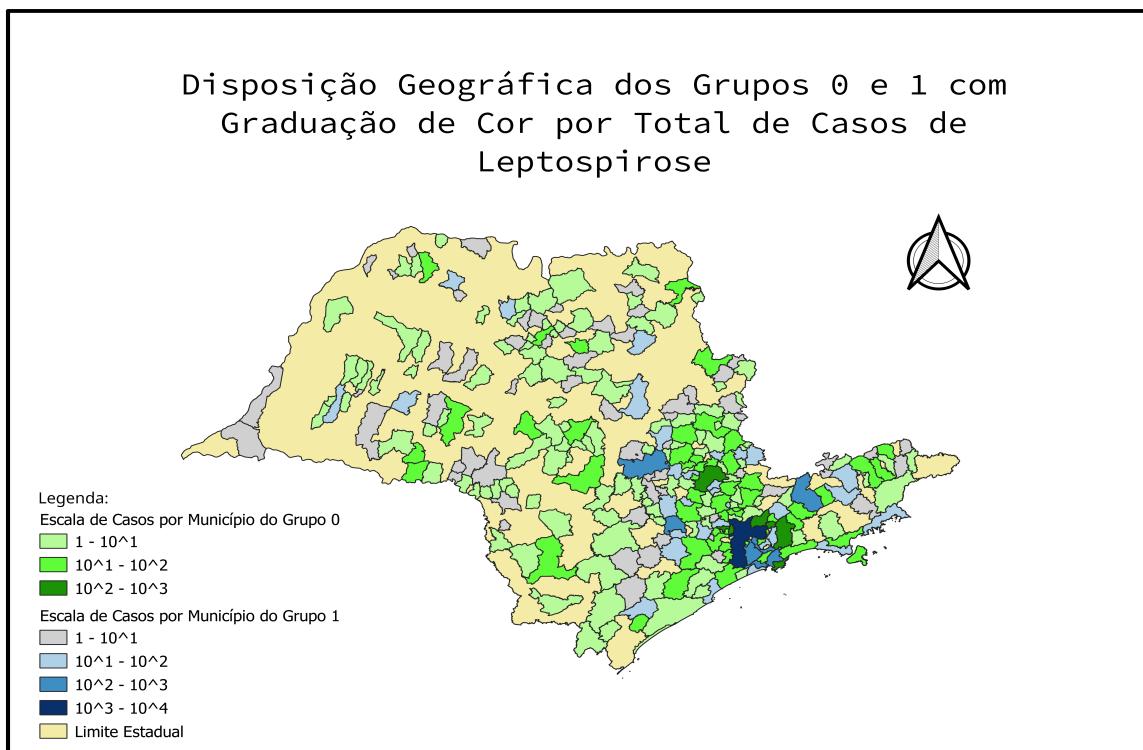


Figura 10 – Disposição Geográfica do Grupo 1 com Graduação de Cor por Total de Casos de Leptospirose. Fonte: próprio autor

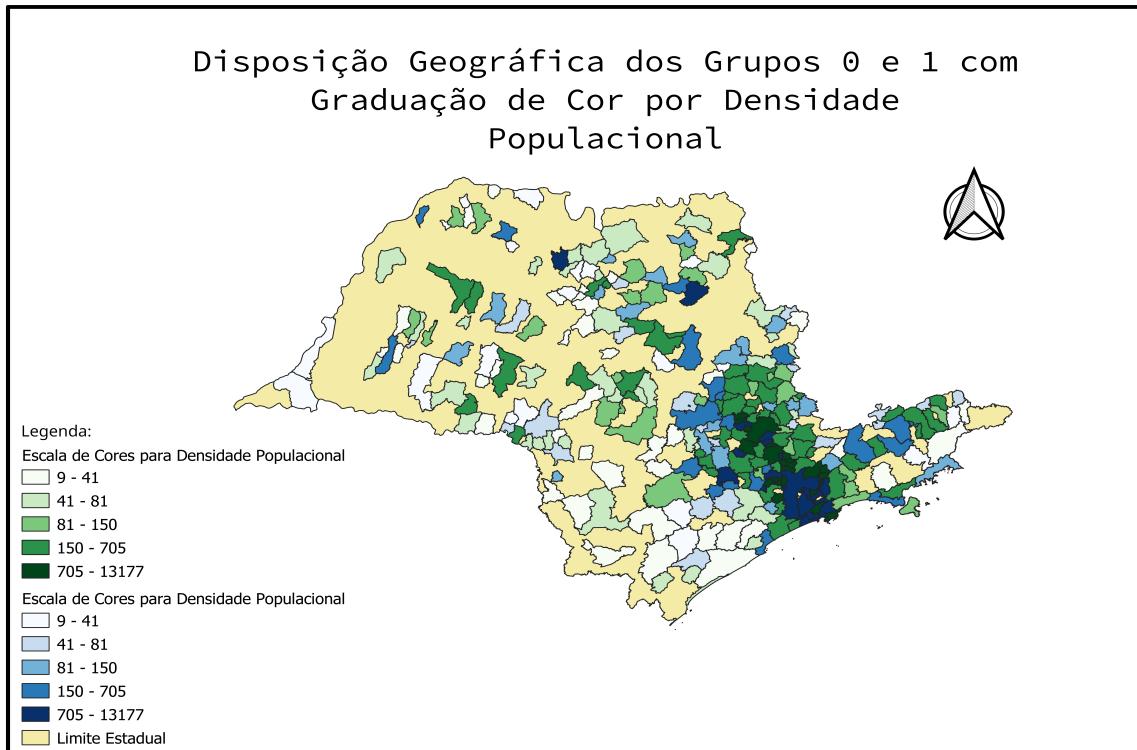


Figura 11 – Disposição Geográfica do Grupo 1 com Graduação de Cor por Densidade Populacional. Fonte: próprio autor

As duas últimas figuras apresentadas (10 e 11) foram visualizações gerada a fim de esclarecer um contexto do estado tanto perante a Leptospirose (Figura 10) quanto perante à concentração populacional nas cidades. Aproximando as duas realidades deseja-se fomentar análises sobre o desenvolvimento histórico do estado e sobre estruturação de esgotos e abastecimento de água, uma vez que a Leptospirose é uma doença intimamente ligada ao ambiente de contaminação ([CLAZER et al., 2015](#)).

6 Discussão

Por meio da comparação dos resultados apresentados na Figura 5 e nas tabelas 1 e 2, é possível traçar um perfil básico dos grupos. De forma geral o *Grupo 0* apresenta as menores médias de todas as características analisadas, enquanto o *Grupo 1* as maiores. Isso significa que os municípios do *Grupo 0* apresentam menor concentração de casos em todos os aspectos registrados pelo DATASUS, a menor média de Total de casos, além de representar municípios com menor Densidade Populacional e o menor IDH-M. Por sua vez, o referenciamento geográfico permite a identificação de uma área de foco em que existem municípios vizinhos de ambos os agrupamentos que apresentam alta densidade populacional e alto índice de ocorrência de leptospirose.

Nota-se que a diferença entre os agrupamentos encontrados são os valores médios das características, mas quando se analisa-as em subgrupos é possível traçar um perfil de contaminação equivalente para ambas as classes. Considerando as características extraídas do DATASUS, temos as maiores médias intragrupo para: contaminação em região urbanizada (subgrupo de características de urbanização), o local de contágio sendo o domicílio e tem-se: “quinta a oitava série incompleta” como escolaridade em comum. Sendo assim, é levantada a hipótese de que o perfil identificado é intensificado conforme a Densidade Populacional dos municípios, uma vez que o *Grupo 0* possui 187 municípios e o *Grupo 1* 104, porém o *Grupo 1* possui 3075 registros confirmados de leptospirose em contraste com 748 registros que o *Grupo 0* carrega. O perfil geral identificado também indica a hipótese de condições ruins de saneamento básico nas moradias, um agravante comum para contaminação (GENOVEZ, 2009), que pode indicar a condição econômica precária dessa parcela populacional. Esse perfil envolve o fenômeno social de evasão no ensino para atuação precoce no mercado de trabalho (ENGELS; DUENHAS, 2020). Quando essa evasão ocorre nos últimos quatro anos, ou antes, do ensino fundamental, como levantado por esta pesquisa em questão e apresentados na Figura 5, o conhecimento dessas pessoas que evadiram é insuficiente para assimilação de como as atitudes podem comprometer a saúde e levar à contaminação de certas doenças (baseado nas ementas do MEC - Ministério da Educação e Cultura - sobre ensino de Ciências Naturais¹ e de Temas Transversais da Saúde²), como a Leptospirose.

¹ <http://basenacionalcomum.mec.gov.br/images/pcn/ciencias.pdf>

² <http://basenacionalcomum.mec.gov.br/images/pcn/saude.pdf>

6.1 Trabalhos Futuros

Pelo fato do desvio padrão ter sido elevado na maioria das características de cada agrupamento (características presentes na Figura 5 e 6 e nas tabelas 1 e 2), sugere-se a adição do índice de Gini (indicador que varia de 0 a 1, sendo que quanto mais próximo de 1 mais desigual economicamente um município se apresenta [Wolffenbuttel \(2004\)](#)) para uma nova análise (disponível no DATASUS), assim como a especificação do tratamento de esgoto do município e abastecimento de água (disponíveis no SNIS³). Devido a baixa quantidade de municípios de São Paulo disponíveis para consulta (293 fornecidos pelo DATASUS / 645 municípios do estado, segundo o IBGE) foi acumulada uma quantidade de casos de Leptospirose de 2007 a 2020, portanto, para a inclusão de novos indicadores é necessário que a série histórica seja repensada e a amostra restringida, uma vez que o SNIS possui dados a partir de 2014.

Como proposta alternativa à adição de colunas à base de dados existente sugere-se a aplicação de outro algoritmo de clusterização que seja menos sensível a ruídos a fim de comparação dos resultados ([ESTER et al., 1996](#)). O *DBSCAN (Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise)*, por exemplo, além de lidar bem com ruídos, também é eficiente para agrupar dados com formas variadas. O *DBSCAN* pode ser aplicado na base de dados com características geográficas por possuir afinidade aos bancos de dados espaciais ([ESTER et al., 1996](#)), o que pode revelar padrões nos municípios que dividem fronteira física e possuem concentração superior a 1000 casos, como: Barueri, Osasco, São Paulo, Guarulhos, Itaquaquecetuba e Mogi das Cruzes, mas que não dividem a mesma classe (conforme o agrupamento do Complete Linkage). Apesar de ser um algoritmo com complexidade computacional elevada ([ESTER et al., 1996](#)), o tamanho da amostra coletada não intensifica esse aspecto negativo.

Quanto à análise da dispersão geográfica que os mapas trazem, sugere-se o complemento das camadas para geração de mais visualizações. Os resultados apresentados evidenciam concentração de municípios com alto índice de contaminação (Figura 10) e tal fato pode ser aprofundado por meio da adição de camadas de redes de esgoto em geometria linear no mapa, por exemplo. Unir esses dados geográficos com as informações sobre tratamento de esgoto pode gerar conhecimento para explicações dos picos regionais de contaminação, como identificação de zonas com esgotos a céu aberto. O mesmo pode ser pensado para criação de visualizações sobre o abastecimento de água, a relação dessas informações, reconstruídas em paisagem digital, podem contribuir com interpretações novas das dispersões dos casos por município.

³ <http://app4.mdr.gov.br/serieHistorica/>

Referências

- AGGARWAL CHARU C E REDDY, C. K. **Clustering de Dados: Algoritmos e Aplicações**, ser. [S.l.]: Chapman and Hall/CRC Data Mining and Knowledge Discovery Series. Chapman e ... , 2013.
- AHMED, S. et al. Perfil de unidades básicas de saúde quanto às ações de rastreamento do câncer do colo do útero no rj. In: **XXIV Congresso Brasileiro de Engenharia Biomédica**. [S.l.: s.n.], 2014. v. 273.
- BALDAN, S. S.; NUNES, E. M.; ANDRADE, M. de. Aspectos epidemiológicos e socioeconômicos relacionados aos casos de óbito por tuberculose no estado de mato grosso do sul. 2018.
- BRITO, M. A. de. Bonita r, beaglehole r, kjellstrom t. epidemiologia básica. são paulo: Grupo editorial nacional; 2010. **Ciência & Saúde Coletiva**, Associação Brasileira de Pós-Graduação em Saúde Coletiva, v. 17, n. 6, p. 1657–1658, 2012.
- CAMPOS, A. C. V. et al. Indicadores socioeconômicos e de saúde da atenção básica nos municípios da região metropolitana de belo horizonte. **Arquivos em Odontologia**, v. 48, n. 1, 2012.
- CARVALHO, D. R.; MILANI, C. S. Pós-processamento em kdd. **Revista de Engenharia e Tecnologia**, v. 5, n. 1, p. Páginas–151, 2013.
- CASANOVA, M. A. **Anatomia de Sistemas de Informação Geográfica**. Mundogeo, 2005. Disponível em: <<http://www-di.inf.puc-rio.br/~casanova//Publications/Books/>>.
- CLAZER, M. et al. Leptospirose e seu aspecto ocupacional: revisão de literatura. **Arq Cienc Vet Zool**, v. 18, n. 3, p. 191–8, 2015.
- ENGELS, T.; DUENHAS, R. A. A lei de aprendizagem: Breve análise entre 2009 e 2019, no brasil e no paraná. **Revista Paranaense de Desenvolvimento-RPD**, v. 41, n. 139, 2020.
- ESTER, M. et al. **A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise**. [S.l.]: AAAI Press, 1996. 226–231 p. (KDD'96).
- EVERITT, B. S.; LANDAU, S.; LEESE, M. Cluster analysis arnold. **A member of the Hodder Headline Group, London**, p. 429–438, 2001.
- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37, Mar. 1996. Disponível em: <<https://ojs.aaai.org/index.php/aimagazine/article/view/1230>>.
- FAYYAD USAMA M E PIATETSKY-SHAPIRO, G. e. S. P. e. o. Descoberta de conhecimento e mineração de dados: Rumo a uma estrutura unificadora. In: . [S.l.: s.n.], 1996. v. 96, p. 82–88.
- FIGUEIREDO, C. M. d. et al. Leptospirose humana no município de belo horizonte, minas gerais, brasil: uma abordagem geográfica. **Revista da Sociedade Brasileira de Medicina Tropical**, SciELO Brasil, v. 34, p. 331–338, 2001.

- FINE, P. et al. John snow's legacy: epidemiology without borders. **The Lancet**, Elsevier, v. 381, n. 9874, p. 1302–1311, 2013.
- FITZ, P. R. **Geoprocessamento sem complicações**. [S.l.]: Oficina de textos, 2018.
- FLOREK, K. et al. Sur la liaison et la division des points d'un ensemble fini. In: **Colloquium mathematicum**. [S.l.: s.n.], 1951. v. 2, n. 3-4, p. 282–285.
- FONSECA, G. H. G. da. Fundamentos de banco de dados. 2020.
- GENOVEZ, M. E. Leptospirose: uma doença de ocorrência além da época das chuvas. **Biológico**, v. 71, n. 1, p. 1–3, 2009.
- HAIR, J. F. et al. **Análise multivariada de dados**. [S.l.]: Bookman editora, 2009.
- HAND, D. J. **Data clustering: Theory, algorithms, and applications by guojun gan, chaoqun ma, jianhong wu**. [S.l.]: Wiley Online Library, 2008.
- JAIN, A. K. Agrupamento de dados: 50 anos além do *K-means*. **Cartas de reconhecimento de padrão**, Elsevier, v. 31, p. 651–666, 2010.
- JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Prentice-Hall, Inc., 1988. Disponível em: <https://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf>.
- JOE, H. W. Agrupamento hierárquico para otimizar uma função objetivo. **Jornal da associação estatística americana**, JSTOR, v. 58, n. 301, p. 236, 1963.
- JOHNSON, S. C. Hierarchical clustering schemes. **Psychometrika**, Springer, v. 32, n. 3, p. 241–254, 1967.
- LUCIA, C. A. S. S. F. D. **Visualização de dados em Sistemas de Informação Geográfica: uma revisão sistemática da literatura**. [S.l.], 2017.
- MACQUEEN, J. Classification and analysis of multivariate observations. In: **5th Berkeley Symp. Math. Statist. Probability**. [S.l.: s.n.], 1967.
- NASCIMENTO, A. **Avaliação de farmácias hospitalares brasileiras utilizando análise de correspondência múltipla**. Tese (Doutorado) — Dissertação de mestrado. Rio de Janeiro: UFRJ, 2011.
- OLIVEIRA, E. F. d. Um sistema para cálculo de rotas de caminho mínimo utilizando pgRouting e dados do openstreetmap. Universidade Federal de Uberlândia, 2021.
- ŘEZANKOVÁ HANA E EVERITT, B. Análise de cluster e dados categóricos. v. 89, n. 3, p. 216–232, 2009.
- ROUSSEEUW, P. J. Silhuetas: uma ajuda gráfica para a interpretação e validação da análise de cluster. **Revista de matemática computacional e aplicada**, Elsevier, v. 20, p. 53–65, 1987.
- SANTOS, E. P. **Mineração de dados aplicada à tuberculose nos municípios do Estado de São Paulo**. Tese (Doutorado) — Universidade de São Paulo, 2020.

- SANTOS, J. A. d. **Algoritmos rápidos para estimativas de densidade hierárquicas e suas aplicações em mineração de dados.** Tese (Doutorado) — Universidade de São Paulo, 2018.
- SANTOS, J. R. R. dos. **Ciência de Dados e Políticas Públicas de Saúde: Exemplos Práticos.** [S.l.], 2020.
- SILVA, I. C. d. **Geoprocessamento e biopolítica: Vigilância espacial por meio do Sistema Único de Saúde.** Dissertação (B.S. thesis), 2019.
- SOUZA, A. P.; ZAIA, J. E. O uso do data mining na promoção de saúde: uma revisão sistemática da literatura. **Atas de Saúde Ambiental-ASA (ISSN 2357-7614)**, v. 3, n. 1, p. 12–21, 2015.
- WOLFFENBUTTEL, A. **O que é? - Índice de Gini.** 2004. Disponível em: <https://www.ipea.gov.br/desafios/index.php?option=com_content&id=2048:catid=28>.
- XU, R.; WUNSCH, D. Survey of clustering algorithms. **IEEE Transactions on neural networks**, Ieee, v. 16, n. 3, p. 645–678, 2005.