INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS
DEPARTAMENTO DE ÁREAS ACADÊMICAS DO CÂMPUS JATAÍ
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

FLÁVIA LOPES SOUSA

KDD APLICADO À ANÁLISE ESPAÇO TEMPORAL DA LEISHMANIOSE VISCERAL NO ESTADO DO PARÁ ENTRE OS ANOS DE 2007 E 2019

Jataí

Março de 2022

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS
DEPARTAMENTO DE ÁREAS ACADÊMICAS DO CÂMPUS JATAÍ
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

FLÁVIA LOPES SOUSA

KDD APLICADO À ANÁLISE ESPAÇO TEMPORAL DA LEISHMANIOSE VISCERAL NO ESTADO DO PARÁ ENTRE OS ANOS DE 2007 E 2019

Trabalho de Conclusão de Curso apresentado ao Instituto Federal de Goiás – Campus Jataí como um dos pré-requisitos necessários para aprovação no Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas.

Jataí

Março de 2022

Ficha catalográfica

FLÁVIA LOPES SOUSA

KDD APLICADO À ANÁLISE ESPAÇO TEMPORAL DA LEISHMANIOSE VISCERAL NO ESTADO DO PARÁ ENTRE OS ANOS DE 2007 E 2019

Trabalho de Conclusão de Curso apresentado ao Instituto Federal de Goiás – Câmpus Jataí como um dos pré-requisitos necessários para aprovação no Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas.

Aprovado em:

BANCA EXAMINADORA

Prof. Me. Roney Lopes Lima, Mestre

Instituto Federal de Goiás

Orientador

Prof. Dr. Aladir Ferreira da Silva Júnior, Doutor

Instituto Federal de Goiás

Banca Examinadora

Prof. Dr. Gustavo de Assis Costa, Doutor

Instituto Federal de Goiás

Banca Examinadora

Dedico este trabalho às vítimas da Leishmaniose Visceral no estado do Pará e a minha saudosa tia avó, Maria Alvina Henrique Pinto.

AGRADECIMENTOS

Deus é maravilhoso e benevolente. A primeira bondade para comigo nesta existência foi me fazer vir ao mundo através de você, mãe. Afinal, o que teria sido de mim sem a sua dedicação? Não poderia deixar de iniciar esta escrita sem deixar registrado o meu agradecimento público: obrigada por tudo, você me ensinou o que é o verdadeiro amor. Também agradeço a toda minha família por todo o apoio e carinho tão essenciais.

"Ajuda-te a ti mesmo, que o céu te ajudará". Deus pede para que eu me ajude, mas em muitos momentos, quando as minhas forças se esgotaram, ele me pegou pela mão e me guiou, colocou pessoas incríveis no meu caminho. E a minha gratidão também é por todas essas pessoas, que tanto fizeram a diferença na minha trajetória e que me inspiraram, e inspiram, pela palavra ou pela ação. Saibam, vocês fortalecem em mim o desejo de gerar impacto positivo na vida do próximo e dar o meu melhor a cada dia.

Quero agradecer de forma geral a todos os profissionais da Instituição (IFG) que trabalharam com dedicação, afinco e seriedade, permitindo que todos os alunos, inclusive eu, tivessem a melhor experiência de ensino, apesar dos obstáculos que tivemos nos últimos anos. Agradeço especificamente a todos os professores do curso de TADS por todas as experiências trocadas, por se empenharem em formar excelentes analistas e desenvolvedores de sistemas, vocês são especiais.

Não poderia deixar de citar três professores que marcaram a minha passagem pela Instituição. Ao professor Gustavo de Assis Costa: obrigada por compartilhar sua experiência trazendo para a sala de aula mais do que conteúdo; obrigada pela disponibilidade e empatia; e obrigada por me ajudar a identificar uma área de interesse que estou transformando em competência, que é a análise de dados. O meu agradecimento também vai ao professor Leizer Fernandes Moraes, um exímio professor que me ajudou a desenvolver um pensamento científico e melhorar a minha escrita durante as aulas das disciplinas de TCC.

Finalmente, quero agradecer ao professor Roney Lopes Lima, um professor excelente, dedicado, competente, que contribuiu muito para o meu desenvolvimento, desde a Iniciação Científica até a finalização deste Trabalho de Conclusão de Curso. Obrigada por toda a atenção e contribuições, sempre será lembrado com carinho.

Encerro com a seguinte constatação: o ensino público de qualidade muda vidas e esta deve ser uma bandeira erguida e defendida por toda a sociedade.

“Tudo está relacionado com todo o resto, mas as coisas próximas estão mais relacionadas do que as coisas distantes.”

WALDO TOBLER

RESUMO

A Leishmaniose Visceral Humana (LVH) é uma doença parasitária grave e negligenciada presente em vários continentes. No Brasil, o estado do Pará abriga os municípios com o maior risco de transmissão das Américas no triênio 2017-2019. Este estudo teve como objetivo identificar áreas de risco no referido estado, o período em que ocorreram e a distribuição da doença no espaço e tempo. Através de uma abordagem de Descoberta de Conhecimento em Bases de Dados realizou-se uma análise exploratória retrospectiva na base de notificações de LVH do SINAN entre os anos de 2007 e 2019. O processo de KDD foi adotado como metodologia de análise dos dados. O índice I de Moran Global foi positivo em todo o período e maior em 2019 (0,45). Os índices I de Moran Local (estatística LISA) indicaram que os municípios com maior risco estavam no nordeste e sudeste do estado. A Estatística de Varredura Espaço-Temporal identificou três agrupamentos circulares, com raio de até 178 km, sendo dois entre 2007 e 2013 no nordeste do estado, e outro entre 2014 e 2019, no sudeste do estado.

Palavras-chave: KDD; *Data Mining*; agrupamento espaço-temporal.

ABSTRACT

The Human Visceral Leishmaniasis (LVH) is a serious and neglected parasitic disease present in several continents. In Brazil, the Para state has the municipalities with the highest transmission risk in the Americas, considering the period 2017-2019. This study aimed to identify risk areas in Para state, the occurrence period and the illness distribution in space and time. Through a Knowledge Discovery in Databases approach a retrospective exploratory analysis was performed on SINAN LVH data between the years 2007 and 2019. The KDD process was used as a data analysis methodology. The Global Moran 's I were positive throughout the period and higher in 2019 (0,45). The Local Moran ’s I (LISA statistic) indicated that the municipalities with higher risk were in the northeast and southeast of state. The spatiotemporal scan statistics found three circular clusters with a radius of up to 178 km (about 110.6 mi), two of them occurred between 2007 and 2013 in the northeast, and the other one occurred between 2014 and 2019, in the southeast.

Keywords: KDD, Data Mining, spatiotemporal clustering.

LISTA DE FIGURAS

LISTA DE EQUAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A Leishmaniose Visceral Humana (LVH) é uma doença parasitária, sistêmica, grave e negligenciada, transmitida por vetor[1] (vetorial), que pode levar à morte em até 90% dos casos quando não tratada. O principal reservatório[2] da doença em peridomicílios[3], no Brasil, é o cão doméstico (leishmaniose canina) (BRASIL *et al.*, 2019). Algumas cidades no centro-sul do estado do Pará, no final da década de 2020, relataram aumento de casos novos da doença, gerando preocupação na população. Em Ourilândia do Norte, para conter o avanço da infecção, dentre outras medidas, um canil[4] foi criado pela administração local para avaliação de cães abandonados.

Na mesma época, alguns municípios próximos estavam lidando com taxas de LVH muito superior ao esperado. Estes acontecimentos motivaram a investigação dos dados de notificações de LVH no estado do Pará, com o objetivo de analisar a dinâmica espacial da doença nos últimos anos e verificar se houve excesso de casos em regiões próximas, caracterizando padrão de agrupamento.

Ao passo que os dados foram obtidos e analisados, foi realizada uma pesquisa para melhor entendimento do ciclo da doença. Foi constatado que no Brasil a doença está espalhada nas cinco regiões brasileiras, atingindo 21 unidades federativas (BRASIL *et al.*, 2019). Em 2019, 97% dos casos das Américas notificados ao SisLeish (Sistema de Informação Regional de Leishmanioses) tiveram origem no território brasileiro, de acordo com o relatório da Organização Pan-Americana da Saúde (OPAS, 2020), sobre leishmanioses.

O mesmo relatório apresentou o cálculo do índice de risco de transmissão da LVH para todas as Américas, no triênio 2017-2019, e estratificou o índice em cinco categorias que vão de risco baixo a muito intenso. Todos os locais com risco de transmissão alto, intenso e muito intenso foram localizados no Brasil. Os dois únicos municípios com risco de transmissão muito intenso, no período, foram localizados no sudeste paraense.

Esta pesquisa também identificou que focos relevantes da doença se desenvolveram,

---

[1] Vetor - Vetor é todo ser vivo capaz de transmitir um agente infectante, de maneira ativa ou passiva. Disponível em: https://pt.wikipedia.org/wiki/Vetor_(epidemiologia). Acesso em: 23 mai. 2022.

[2] Reservatório - Reservatório é um hospedeiro de outra espécie, que alberga o agente etiológico de determinada doença. Disponível em: https://pt.wikipedia.org/wiki/Reservat%C3%B3rio_(medicina). Acesso em: 23 mai. 2022.

[3] Peridomicílios - Peridomicílio é definido como a área externa de uma residência, em um raio não superior a cem metros. Disponível em: https://pt.wikipedia.org/wiki/Peridomic%C3%ADlio. Acesso em 23 mai. 2022.

[4] CANIL MUNICIPAL! - Prefeitura de Ourilândia do Norte – Facebook. 16 jul. 2020. Disponível em: https://www.facebook.com/pmonoficial/videos/canil-municipal/2680476282053321/. Acesso em 4 mar. 2022.

entre 1999 e 2015, no oeste de São Paulo (CARDIM *et al.*, 2016), Minas Gerais (SILVA *et al.*, 2020), Maranhão (FURTADO *et al.*, 2015) e Sergipe (ARAÚJO, 2017), e na região norte, Tocantins (FONTOURA *et al.*, 2016) e Pará (OLIVEIRA *et al.*, 2019) (CARVALHO *et al.*, 2019). Entre 2004 e 2014, 20% dos casos de LVH no Brasil foram notificados em cinco cidades: Fortaleza (Ceará), Campo Grande (Mato Grosso do Sul), Araguaína (Tocantins), Belo Horizonte (Minas Gerais) e Teresina (Piauí) (MACHADO *et al.*, 2019). No Pará, os casos de LVH têm aumentado nos últimos anos, considerando o período de estudo 2007-2019, chegando ao ápice no último triênio, 2017-2019.

Quando em um ambiente se encontram as condições ideais (vetores, parasitas, reservatórios e pessoas vulneráveis), a doença tende a formar agrupamentos locais que, por sua vez, tendem a ser duradouros, se não houver plano eficiente de intervenção. E, finalmente, a doença pode se expandir para locais próximos e formar novos focos.

Agrupamentos espaço-temporais da doença indicam que uma determinada região possui fatores que estão facilitando a permanência da infecção. Os serviços de vigilância epidemiológica, principalmente em localidades com presença regular da doença, podem monitorar os casos de LVH a fim de identificar áreas com casos excessivos, que fogem do esperado, e planejar ações para conter o avanço da doença.

A identificação desses padrões espaciais e temporais são essenciais para entender como a doença se propaga em um território. Vários trabalhos investigaram a distribuição espacial e temporal da LVH, e algumas das várias técnicas utilizadas para este fim foram: os índices de autocorrelação espacial global e local; estimativa de densidade de kernel; varredura espacial, temporal e espaço-temporal. Alguns destes trabalhos são relatados no subcapítulo 2.2, Trabalhos Relacionados.

Como demonstrado, a LVH é um problema de saúde pública e apresenta focos em todo o país. Quando estes focos não são controlados, eles se expandem no território e podem apresentar duração maior, como sugeriram alguns dos trabalhos analisados.

Através de uma abordagem de Descoberta de Conhecimento em Bases de Dados (KDD) e técnicas estatísticas, os dados do SINAN (Sistema de Informação de Agravos de Notificação), que é um sistema desenvolvido e administrado pelo DATASUS, (Departamento de Informática do Sistema Único de Saúde), foram explorados com o objetivo de identificar padrões espaciais e temporais na distribuição da doença no território paraense no período 2007-2019. O KDD é um processo organizado, iterativo e interativo que possui uma série de etapas. A Mineração de Dados, a *Data Mining* (DM), é o núcleo do processo e extrai informações dos dados através de

seus métodos.

Este trabalho adota como metodologia o processo do KDD, que é composto de três principais etapas: pré-processamento, mineração dos dados e pós-processamento. Os métodos de mineração utilizados são estatísticos e têm a função de buscar padrões nos dados considerando o espaço (I de Moran global e local, LISA) e espaço-tempo (Estatística de Varredura Espaço-Temporal, SaTScan).

O Capítulo II (Fundamentação Teórica) apresenta em mais detalhes as teorias que embasaram este trabalho, assim como mais detalhes sobre a Leishmaniose Visceral, LV. O Capítulo III (Metodologia) descreve o processo de KDD, a área de estudo, fonte e estrutura dos dados. O Capítulo IV (Implementação) discorre sobre o processo de análise dos dados e o Capítulo V (Testes e Avaliações), apresenta uma compilação dos resultados obtidos na fase de Implementação. Finalmente, o Capítulo VI (Conclusões e Trabalhos Futuros) apresenta as conclusões sobre a pesquisa e indica pontos de extensão.

2 FUNDAMENTAÇÃO TEÓRICA

O processo de Descoberta de Conhecimento em Bases de Dados, ou *KDD* (*Knowledge Discovery in Databases*), é composto de várias etapas operacionais que são guiadas pelos objetivos definidos pelos atores do processo (especialistas de domínio e especialistas de *KDD*) para a aplicação. É comum o termo Mineração de Dados, ou DM (*Data Mining*), ser utilizado como sinônimo de KDD, mas a Mineração de Dados é uma das etapas e núcleo do processo. A etapa de DM envolve diversas técnicas baseadas em teorias de áreas como a Estatística e Inteligência Artificial, dentre outras.

Neste capítulo, são encontrados mais detalhes sobre a LVH, tema de investigação, e apresentados os referenciais teóricos sobre o *KDD*, DM, e Mineração de Dados Espaço-Temporais, ou STDM (*Spatiotemporal Data Mining*). Durante a exploração de dados espaço-temporais é necessário considerar que os dados podem não ser estacionários, ou seja, pode existir dependência espacial. Esta pode ser verificada utilizando-se técnicas estatísticas de autocorrelação espacial. Este capítulo também apresenta estes conceitos e as técnicas estatísticas usadas para detecção de agrupamentos espaciais que são o Índice de Moran Global e Local, e Estatística de Varredura Espaço Temporal, que serão utilizadas como método de mineração de dados para identificação de agrupamentos no espaço-tempo.

2.1 LEISHMANIOSE VISCERAL HUMANA

A Leishmaniose Visceral, LV, é uma doença parasitária, sistêmica, negligenciada, grave e de distribuição global que atinge principalmente a África, a Ásia e as Américas. É endêmica[5] em 13 países americanos e do total de casos registrados nas Américas, em 2019, 97% (2.529) foram notificados no Brasil. Todas as localidades com risco de transmissão muito intenso, intenso e alto no período 2017-2019, registrados nos continentes americanos, estão localizadas em território brasileiro. Sua transmissão é vetorial e integra, assim como as leishmanioses cutânea e mucocutânea, o grupo de doenças infecciosas negligenciadas. Populações empobrecidas, com pouco, ou nenhum, acesso a serviços de saúde são as mais vulneráveis (OPAS, 2020).

A LV é causada por parasitas do gênero *Leishmania*, pertencentes à família

[5] Endêmica – uma doença é endêmica quando está restrita à uma região geográfica de maneira contínua, normalmente essas regiões oferecem condições ideais para a manutenção do seu ciclo de transmissão (ex.: clima, presença de vetores). Disponível em: https://pt.wikipedia.org/wiki/Endemia. Acesso em 23 mai. 2022.

Trypanosomatidae, estes são digenéticos[6], apresentando duas formas evolutivas. No Brasil, a espécie do parasita mais comum é a *Leishmania infantum chagasi*. Os vetores são as fêmeas de flebotomíneos, insetos da subfamília *Phlebotominae*, e no Brasil, as espécies mais comuns são a *Lutzomyia longipalpis*, seguida da *Lutzomyia cruzi* (BRASIL, 2019). Estes são flebotomíneos, popularmente chamados de mosquito palha na maioria das regiões do Brasil.

No ambiente silvestre, a raposa do campo, *Cerdocyon thous*, foi identificada pela primeira vez no estado do Pará como um reservatório primário do parasita *Leishmania infantum chagasi*. Este animal convive harmonicamente com o parasita sem apresentar sintomas clínicos. O cão doméstico, *Canis familiaris*, é o hospedeiro secundário acidental e já apresenta vários sintomas clínicos como emagrecimento, onicogrifose[7], ceratite [8]periorbital, descamação e lesões ulcerativas de pele (SILVEIRA *et al.*, 2016).

O ciclo de transmissão envolve um mosquito fêmea, hematófaga[9], que ao picar um animal infectado passa a hospedar o parasita no seu intestino. Posteriormente, os parasitas migram para a saliva do mosquito que contamina os novos animais através de sua picada. O tempo de incubação do parasita varia dependendo do hospedeiro, sendo no ser humano entre 10 dias e 24 meses, com média entre 2 e 6 meses. Já no cão, o tempo de incubação é de 3 meses a alguns anos, com média de 3 a 7 meses. Apenas uma pequena parcela de humanos manifesta sintomas da doença, dependendo do seu sistema imunológico. É sabido que crianças e idosos são mais vulneráveis, também pessoas com a imunidade comprometida ou subnutridas. A LVH é uma forma menos frequente da doença em comparação à *Leishmaniose Tegumentar Americana*, a forma cutânea, e sua variante mucocutânea, que atinge as mucosas, entretanto mais grave podendo levar a óbito em mais de 90% dos casos quando não tratados (BRASIL, 2019).

A LV é uma doença que possui uma dinâmica complexa, diversas variáveis podem influenciar no aumento de casos em uma região (fatores socioambientais, socioeconômicos, climáticos). Alguns trabalhos associam a migração, intra e interestadual, como um dos fatores

---

[6] Digenéticos - Os Parasitas do gênero Leishmania são digenéticos (heteroxenos) e apresentam em seu ciclo de vida apenas duas formas evolutivas: a forma promastigota, que é flagelada e extracelular, e a forma amastigota, que é intracelular e sem movimentos. Disponível em: http://www.dbbm.fiocruz.br/tropical/leishman/leishext/html/morfologia.htm. Acesso em: 23 mai. 2022.

[7] Onicogrifose - Onicogrifose é uma distrofia ungueal também chamada de "unhas de chifres de carneiro" ou "unhas em garra". Disponível em: https://pt.wikipedia.org/wiki/Onicogrifose. Acesso em: 23 mai. 2022.

[8] Ceratite - Ceratite ou queratite é uma inflamação na córnea, a camada transparente que protege os olhos. Pode ser causada por secura, lesão física ou química, vírus, bactérias, amebas, fungos ou vermes. Disponível em: https://pt.wikipedia.org/wiki/Ceratite. Acesso em: 23 mai. 2022.

[9] Hematófaga - Hematófago é um grupo de animais ou parasitas que se alimentam de sangue. Disponível em: https://pt.wikipedia.org/wiki/Hemat%C3%B3fago. Acesso em: 23 mai. 2022.

de expansão da doença que pode contribuir para que áreas indenes se tornem focos de LVH, pois uma área livre do parasita, mas com a presença de vetores, precisa apenas de fontes (reservatórios) do parasita para que a doença seja espalhada.

É o que pode ter acontecido, de acordo com Silveira *et al.* (2016), em algumas cidades do sul do Pará – Redenção e Conceição do Araguaia, na primeira década deste século. A hipótese é a de que cães infectados foram introduzidos na região trazidos do estado vizinho, Tocantins, uma área endêmica, visto que há intenso fluxo migratório oriundo das cidades de Palmas e Araguaína, no Tocantins.

O tempo para apresentação dos primeiros sintomas na pessoa e no animal doméstico (cão), a partir do momento da infecção, pode variar, em média, entre 2 a 7 meses (BRASIL, 2019). Considerando que tanto seres humanos quanto animais infectados podem trafegar de uma região para outra, e que o vetor está presente em quase todo território nacional, pode-se concluir que a Leishmaniose é uma doença infecciosa parasitária que varia no espaço e no tempo de acordo com três principais fatores: dinâmica de movimentação de reservatórios e vetores (este em menor escala); existência de vetores da LV no ambiente e densidade vetorial; e fatores socioambientais, socioeconômicos e climáticos.

A LVH é uma doença de notificação compulsória, com frequência semanal, conforme definido pela portaria n° 204, de 17 de fevereiro de 2016, do Ministério da Saúde (BRASIL, 2016), ou seja, é obrigatória a comunicação ao Ministério da Saúde.

2.2 TRABALHOS RELACIONADOS

Hao *et al.* (2021) investigaram o padrão de distribuição temporal e espacial da *Leishmaniose Visceral Zoonótica* do tipo montanha na China, entre os anos de 2015 e 2019. A autocorrelação espacial foi medida utilizando os índices globais I de Moran e Getis-Ord no *software* ArcGis. Uma análise retrospectiva em busca de agrupamentos espaço-temporais foi realizada através da Estatística de Varredura Espaço-Temporal (SaTScan), utilizando um modelo de Poisson.

Zheng *et al.* (2020) analisaram a LVH, entre os anos 2004 e 2018, em algumas províncias da China com o objetivo de encontrar padrões espaço-temporais e identificar áreas de risco de transmissão. Além disso, verificaram a associação de casos de LVH com fatores meteorológicos. Uma das técnicas utilizadas para exploração de agrupamentos espaciais foi o índice I de Moran.

Cardim *et al.* (2016) avaliaram no espaço e espaço-tempo a distribuição de casos de LVH no estado de São Paulo, entre 1999 e 2013, através de um estudo descritivo e ecológico, utilizando ferramentas de análise espacial Kernel e razão Kernel e estatística de varredura espaço-temporal para a detecção de agregados da doença. Foram detectados dois aglomerados, ambos com duração de 6 anos, e constatado que a doença se expandiu do oeste para o leste do estado, no sentido de uma rodovia, apresentando uma possível associação entre fluxos migratórios e de mercadoria com o espalhamento da doença.

Por meio de um estudo ecológico, Silva *et al.* (2020) analisaram os dados de LVH da Região Metropolitana de Belo Horizonte, Minas Gerais (RMBH), no período de 2006 a 2017. A área de estudo compreende os 34 municípios da RMBH e 16 municípios que fazem fronteira com a região. O objetivo foi analisar os padrões espaciais e espaço-temporais de ocorrência de LVH e identificar áreas de risco prioritárias para vigilância e controle na região. Foram utilizados indicadores de autocorrelação espacial global (I de Moran) e local (LISA) para identificar áreas prioritárias. A detecção de agregados espaço-temporais foi realizada utilizando a Estatística de Varredura Espaço-Temporal, com modelo de probabilidade Poisson Discreto, implementada no *software* SaTScan. Foi identificada alta concentração de casos de LVH no núcleo metropolitano e dez municípios foram considerados de alto risco.

Araújo (2017) analisou espacialmente os casos de LVH no estado de Sergipe, entre os anos 2010 e 2015, através de um estudo ecológico, descritivo e retrospectivo, utilizando métodos estatísticos e de visualização como mapa coroplético (distribuição do total de casos), mapa do estimador de densidade de kernel (visualização de pontos quentes), cálculo de índice de autocorrelação espacial global (I de Moran) e local (LISA). Foram encontrados indícios de expansão da doença em municípios circunvizinhos a Aracaju, municípios com notificação persistente da doença e focos em outras regiões do estado.

Fontoura et al. (2016) estudaram a distribuição da LVH no estado do Tocantins, entre os anos de 2008 e 2011, através de um estudo ecológico e exploratório utilizando as mesmas técnicas do trabalho anterior. Identificou a microrregião do Bico do Papagaio, fronteira entre os estados de Tocantins, Maranhão e Pará, como a de maior intensidade na transmissão da doença no estado do Tocantins no ano de 2008. Em 2011, a microrregião com maior intensidade foi a de Araguaína. Os autores atribuíram essa variação espacial à migração entre as microrregiões tocantinenses. Os três municípios mais afetados pela doença no período de estudo foram Araguatins, Araguaína e Ananás, ambos no estado do Tocantins, todos próximos à fronteira com o sudeste do Pará.

Através de técnicas estatísticas, Furtado *et al.* (2015) analisaram a distribuição temporal e espacial da LVH no estado do Maranhão para cada biênio, entre os anos de 2000 e 2009, por meio de um estudo ecológico retrospectivo. Como resultado foram elaborados mapas temáticos e identificadas áreas de risco nas dezoito unidades regionais de saúde do estado. Foram utilizados um modelo Bayesiano espaço-temporal, para identificação de áreas de risco, e o método MCMC (Markov Chain Monte Carlo) para a estimativa dos parâmetros do modelo. Através dos mapas, percebe-se que as unidades regionais de Açailândia e Imperatriz, ambas no estado do Maranhão, fronteiriças com Pará e Tocantins, têm o risco relativo sempre aumentado a cada biênio.

Estudos descritivos realizados no município de Marabá, sudeste paraense, próximo às fronteiras com Maranhão e Tocantins, indicaram um grande aumento da quantidade de casos a partir de 2015. Marabá registrou números baixos e estáveis até 2014, a partir de 2015 enfrentou uma alta taxa de incidência de LVH (CARVALHO *et al.*, 2019, OLIVEIRA *et al.*, 2019).

2.3 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

O KDD é o processo organizado, iterativo e interativo de extrair informações válidas, novas, úteis e compreensíveis de bases de dados grandes e complexas (MAIMON; ROKACH, 2010).

Este possui várias etapas desde o entendimento do negócio e definição dos objetivos até a avaliação e apresentação dos resultados, sendo o *Data Mining* uma das etapas e núcleo do processo, que consiste no emprego de técnicas inteligentes e automatizadas para obtenção de informações válidas, novas, úteis e compreensíveis a partir dos dados. Enquanto alguns veem a Mineração de Dados como uma etapa de todo o processo, outros consideram o termo como sinônimo de KDD. Por esta razão, Han e Kamber (2012, p. 8) definem *Data Mining* como "[...] o processo de descobrir padrões e conhecimentos interessantes de grandes quantidades de dados".

O processo de Descoberta de Conhecimento em Bases de Dados pode ser separado em três principais etapas operacionais: i) pré-processamento, que envolve funções relacionadas à captação, organização, tratamento e preparação dos dados para a fase posterior; ii) mineração dos dados, que tem a função de extrair conhecimento dos dados e; iii) pós-processamento, que envolve funções relacionadas a avaliar o conhecimento adquirido, organizá-lo e apresentá-lo.

Cada uma das três etapas é composta por várias operações primárias. De acordo com Goldschmidt e Passos (2005, p. 16), uma "[...] Operação de KDD se refere a qualquer função das etapas operacionais de KDD. Uma operação de KDD é, portanto, a especificação, no nível lógico, de uma função de KDD". Uma operação pertencente à etapa de mineração dos dados recebe um nome especial: tarefa de KDD.

Para os autores, uma operação/tarefa primária é aquela que não pode ser desmembrada em outras operações/tarefas, enquanto uma operação/tarefa complexa pode ser desmembrada em duas ou mais operações/tarefas primárias. Os métodos de KDD são implementações específicas das operações e estas implementações são fundamentadas pelas técnicas de KDD. Uma "[...] Técnica de KDD se refere a qualquer teoria que possa fundamentar a implementação de um método de KDD" (GOLDSCHMIDT; PASSOS, 2005).

Baseando-se na definição anterior, pode-se citar a Teoria de Redes Neurais, do campo da Inteligência Artificial, como uma técnica de KDD, pois subsidiou a implementação do método *Support Vector Machine* (Máquina de Vetores de Suporte - SVM), que é aplicado na tarefa de classificação (GOLDSCHMIDT; PASSOS, 2005).

A fase de Transformação, por exemplo, que está inserida na etapa de pré-processamento, possui algumas funções relacionadas à redução de dados como: redução vertical (seleção de subconjunto de atributos), redução horizontal (seleção de subconjunto de amostras) e redução de valores (diminuição de valores distintos em um atributo). Cada uma destas funções, ou operações primárias, é executada por seus métodos específicos. A técnica de Algoritmos Genéticos é uma abordagem de otimização que pode ser utilizada como método da operação de redução vertical dos dados, onde o algoritmo a ser utilizado na fase de mineração pode ser utilizado como sua função de avaliação (GOLDSCHMIDT; PASSOS, 2005).

Até este ponto foram expostas as definições dos termos KDD e Mineração de Dados, onde o primeiro se refere a todo o processo, enquanto o segundo é o núcleo deste. Também foram esclarecidos os termos Operações e Tarefas de KDD, que podem ser simples ou compostas. Foi apresentada a definição de Métodos de KDD como implementações das Operações/Tarefas, os quais são fundamentados por Técnicas de KDD, que são teorias de campos diversos como Inteligência Artificial e Estatística.

O processo de KDD resumido nas três etapas anteriores (pré-processamento, mineração e pós-processamento) é uma representação abstrata, onde cada uma destas é composta por outras etapas mais especializadas. Desse modo, Maimon e Rokach (2010) abstraem o processo em nove etapas, enquanto Han e Kamber (2012) desmembram o processo em sete etapas, que

são: 1) Limpeza; 2) Integração; 3) Seleção; 4) Transformação; 5) Mineração; 6) Avaliação de Padrões e; 7) Apresentação do Conhecimento.

As diferenças entre as descrições do processo nas duas obras são mínimas e estão mais relacionadas à ordenação das tarefas. A primeira obra considera a fase de entendimento do problema e definição dos objetivos como uma etapa, sendo esta a primeira, e divide a etapa de mineração dos dados em três partes: i) escolha da tarefa; ii) escolha do algoritmo e; iii) aplicação do algoritmo. Outra diferença é que as fases de seleção e integração, na primeira obra, são reunidas em uma única etapa que é anterior à fase de limpeza.

A organização das etapas do KDD, proposta por Maimon e Rokach (2010), é mais detalhada e resume melhor o fluxo natural que foi adotado neste trabalho. Ela foi adotada como metodologia e mais detalhes do processo são expostos no "Capítulo III - Metodologia" desta pesquisa.

2.4 MINERAÇÃO DE DADOS

Para Han e Kamber (2012, p. 1-5), o surgimento e desenvolvimento da Mineração de Dados estão fortemente relacionados com a evolução da tecnologia da informação nas últimas décadas. O crescimento explosivo no volume de dados disponíveis nos últimos anos é resultado da informatização da sociedade e rápido desenvolvimento de ferramentas poderosas para coleta e armazenamento de dados.

Vários fatores impulsionaram o aumento dos dados gerados, armazenados e disponíveis para análise. Um deles foi o desenvolvimento de tecnologias de bancos de dados a partir da década de 1970, que evoluíram de modelos hierárquicos e em rede para o modelo relacional, fomentando avanços em termos de otimização de consultas a bancos de dados e gerenciamento de transações. A partir da década de 1980, as pesquisas em bancos de dados avançados se intensificaram, sendo criados modelos novos de dados e soluções em termos de armazenamento e gerenciamento de dados complexos (espaciais, temporais, multimídia, etc.) e de bases de dados muito grandes (HAN; KAMBER, 2012).

O advento da *internet* resultou na criação de bancos de dados baseados na Web (XML e Web Semântica) que possibilitou o desenvolvimento da computação em nuvem, de *streams* de dados e incentivou pesquisas em processamento paralelo de dados (HAN; KAMBER, 2012).

O barateamento de *hardware* para armazenamento de dados, desenvolvimento da *internet* e *intranet*, popularização do uso de sensores, IoT (*Internet of Things*), entre outros, são

fatores que impulsionaram e continuam a impulsionar a produção e disponibilidade de dados. Foi neste cenário que a Mineração de Dados começou a se desenvolver a partir da década de 1980, pois o aumento no volume e variedade dos dados criou a necessidade de técnicas mais apuradas para lidar com os novos desafios (HAN; KAMBER, 2012).

Em suma, estes avanços tecnológicos, além de aumentarem a quantidade de dados armazenados, também incitaram a pesquisa, desenvolvimento e aperfeiçoamento de técnicas para extração de conhecimento de bases de dados variados e em maior volume.

De acordo com Tan *et al.* (2014, p. 6), a mineração de dados foi criada como área de estudo interdisciplinar, que reúne técnicas já existentes de outras disciplinas, estuda, pesquisa e desenvolve novas técnicas de descoberta de conhecimento para enfrentar os desafios de dados. Para estes autores, a Mineração de Dados é uma área interseccional às disciplinas de Estatística, Inteligência Artificial, Aprendizagem de Máquina e Reconhecimento de Padrões, e é apoiada por tecnologias de bancos de dados, computação paralela e computação distribuída. A Mineração de Dados também incorporou ideias de outras áreas como otimização, computação evolutiva, teoria da informação, processamento de sinais, visualização e recuperação de informação.

Maimon e Rokach (2010, p. 5-6) apresentaram a taxonomia de métodos de *data mining* dividindo-os em duas principais categorias: orientados à verificação, onde o sistema verifica a hipótese do usuário (especialista), utilizando técnicas tradicionais da estatística, como testes de hipóteses e análise de variância; e orientados à descoberta, onde o sistema, de forma autônoma, encontra novas regras e padrões. Os métodos orientados à descoberta são constituídos de métodos preditivos (aprendizagem supervisionada) e descritivos (aprendizagem não supervisionada). A seguir são apresentados mais detalhes sobre as tarefas preditivas e descritivas.

2.4.1 Aprendizado Supervisionado: Tarefas Preditivas

Segundo Tan *et al.* (2014, p. 7-8), "Modelagem Preditiva se refere à tarefa de construir um modelo para a variável alvo (dependente) como uma função das variáveis explicativas (independentes)". Ainda para os autores há dois principais tipos de tarefas nesta categoria: as tarefas de classificação, onde a variável-alvo é discreta (ex.: prever o tipo de uma flor: Setosa, Versicolour ou Virginica), e de regressão, onde a variável-alvo é contínua (ex.: prever o preço de uma casa).

Um método preditivo cria um modelo que aprende os padrões em um conjunto de amostras de dados rotulados durante a etapa de treinamento, e em seguida este modelo é exposto a uma outra amostra, ou conjunto de amostras rotuladas, até que se obtenha uma taxa de acerto satisfatória (etapa de teste), resultando em um modelo que será capaz de prever valores do atributo alvo em dados novos e não rotulados. Por essa razão, são chamados de métodos supervisionados.

2.4.2 Aprendizado Não Supervisionado: Tarefas Descritivas

Para Tan *et al.* (2014, p. 7-11), o objetivo das tarefas descritivas é "[...] derivar padrões (correlações, tendências, agrupamentos, trajetórias e anomalias) que resumem as relações subjacentes nos dados".

Os métodos descritivos exploram os dados e buscam entender como estão relacionados, procurando padrões e associações entre amostras e atributos. A Análise de Agrupamentos (*Clustering Analysis*) é a principal tarefa desta categoria. Na próxima seção mais detalhes são apresentados acerca desta temática.

2.4.3 Análise de Agrupamentos

Para Han e Kamber (2012, p. 444), "A análise de cluster [...] é o processo de particionar um conjunto de objetos de dados (ou observações) em subconjuntos. Cada subconjunto é um cluster, de modo que os objetos em um cluster são semelhantes entre si, mas diferentes dos objetos em outros clusters". Em suma, o objetivo da análise de agrupamento é maximizar a similaridade *intracluster* e minimizar a similaridade *intercluster*, utilizando uma medida de similaridade/dissimilaridade adequada para o contexto.

Além de segmentar os dados em grupos semelhantes entre si, esta tarefa pode ser utilizada para detecção de *outliers* e em conjunto com outras tarefas de mineração de dados. Um exemplo seria a aplicação de um algoritmo de *clustering* para definir classes em dados não rotulados e, posteriormente, treinar um modelo de classificação, a partir das classes recém descobertas. Também utilizar técnicas de sumarização nos grupos encontrados, dentre outras possibilidades.

Han e Kamber (2012) ainda destacam que os métodos de *clustering* podem ser comparados de acordo com quatro aspectos ortogonais, que são: i) o critério de

particionamento: os dados podem ser separados em grupos que estão no mesmo nível conceitual ou ter relacionamentos hierárquicos entre si; ii) separação de clusters: os grupos podem ser exclusivos, quando um objeto é exclusivo de um grupo, ou não exclusivos, quando um objeto pode pertencer a mais de um grupo; iii) medida de similaridade: os métodos de agrupamento podem utilizar medidas de similaridade baseadas em distância, que pode ser definida em um espaço euclidiano, ou baseadas em densidade e contiguidade e; iv) espaço de agrupamento: o método pode considerar todo o espaço de dados ou trabalhar em subespaços.

Em relação às categorias de métodos de *clustering*, os autores apresentam quatro principais, que são de métodos baseados em: i) particionamento: grupos não hierárquicos, exclusivos, com similaridade baseada em distância; ii) hierarquia: grupos hierárquicos, não exclusivos, a similaridade pode ser baseada em distância, densidade ou contiguidade, podem considerar todo o espaço ou subespaços; iii) densidade: podem ser hierárquicos ou não, tipicamente exclusivos, a similaridade é baseada em densidade, podem considerar todo o espaço ou subespaços e; iv) grade: a principal característica é que divide o espaço em células, formando uma grade, permitindo um processamento rápido. Pode ser integrado com outros métodos de *clustering*.

Alguns métodos de *clustering* podem apresentar características de mais de uma destas categorias. Os desafios de dados podem exigir que algumas técnicas sejam mescladas para que o objetivo seja alcançado.

2.5 MINERAÇÃO DE DADOS ESPAÇO-TEMPORAIS

Devido à evolução tecnológica e ampla disseminação de dispositivos que coletam dados e os associam com o tempo e posição espacial, sejam dispositivos móveis ou não, surgiu um novo desafio que é a análise de vastas bases de dados espaço-temporais. A Mineração de Dados Espaço-Temporais, ou STDM (*Spatiotemporal Data Mining*) tem como objetivo desenvolver e aperfeiçoar métodos eficientes para extração de conhecimento destes dados.

Quando estes dados são gerados por um dispositivo fixo, que coleta informações periodicamente, temos séries espaço-temporais (ex.: sensores de estações meteorológicas). Se esses dispositivos são móveis, temos dados representando trajetórias. Ao considerar a dimensão espacial, estes dados podem ser pontuais, com o local exato de cada evento, ou agregados por área. Este último é um tipo de dado espaço-temporal muito comum no qual eventos são agregados para representar uma área (ex.: total de casos e taxa de incidência por setor

censitário). Desse modo, a maioria dos dados de saúde do SUS são agregados por área.

Os métodos de STDM são desenvolvidos de acordo com o tipo de dado e objetivo. Muitos dos métodos tradicionais da mineração de dados foram adaptados para serem aplicados nestes tipos de dados. Em relação à tarefa de agrupamento, pode-se citar algoritmos como DBSCAN (BIRANT; KUT, 2007), KNN (JACQUEZ, 1996), dentre outros, que foram adaptados para lidar com dados espaço-temporais.

Cheng *et al.* (2014) realizaram uma revisão sobre STDM acerca de três principais tarefas espaço-temporais, que são: i) modelagem e predição; ii) agrupamento e; iii) e visualização. Os autores apontam uma característica importante deste tipo de dado que é a dependência espacial e temporal. Este fenômeno é a tendência de objetos próximos serem mais semelhantes tanto no espaço quanto no tempo. Por exemplo, em relação a eventos epidemiológicos, se há um surto de uma doença em um dado local, é mais provável que locais próximos possuam taxas de adoecimento semelhantes do que locais mais distantes, e é mais provável que esses eventos ocorram em tempos próximos nas duas localidades.

Ainda destacam que as dependências espacial e temporal violam a suposição de estacionariedade nos dados que métodos estatísticos tradicionais consideram. Logo, tais dependências são características que devem ser consideradas durante o desenvolvimento de métodos de extração de conhecimento para este tipo de dados. A dependência pode ser testada, através de análise de autocorrelação, tanto no tempo quanto no espaço, através de técnicas específicas.

Na seção 2.6, o conceito de autocorrelação espacial é apresentado em mais detalhes e os índices de autocorrelação espacial de Moran, global e local, são descritos. Estes índices foram utilizados na etapa exploratória para identificação de tendência e visualização de agrupamentos espaciais nas taxas médias anuais de LVH nos municípios paraenses.

2.5.1 Agrupamento Espaço-Temporal

O agrupamento espaço-temporal é uma tarefa importante em STDM, cujo objetivo é procurar por associações, padrões implícitos nos dados, sempre buscando maximizar a semelhança entre objetos no mesmo *cluster* e minimizar a semelhança entre objetos de diferentes *clusters*. Os resultados desta tarefa podem embasar hipóteses a serem verificadas posteriormente. O referido agrupamento tem sido amplamente utilizado para extrair padrões de bases de dados espaciais com aplicações em várias áreas como estudos epidêmicos (CHENG *et*

al., 2014).

A similaridade entre os dados espaço-temporais pode ser calculada considerando três domínios: i) temático, composto pelos valores de atributos de cada objeto; ii) espacial, que define a localização de cada objeto e; iii) temporal, que associa o tempo ao objeto. Durante a evolução dos algoritmos de *clustering* foram criados métodos que agrupam dados de acordo com o domínio temático (ex.: K-means), com o domínio espacial (ex.: DBSCAN e BIRCH) e ambos (temático-espacial, temático-temporal). Mas, poucos algoritmos agrupam os dados usando os três domínios simultaneamente, ou seja, considerando objetos semelhantes que estão próximos no espaço e no tempo. A Estatística de Varredura Espaço-Temporal é uma das técnicas de *clustering* em STDM que se destaca (CHENG *et al*., 2014).

2.5.2 Estatística de Varredura Espaço-Temporal

A Estatística de Varredura Espaço-Temporal é uma extensão da varredura puramente espacial. Nesta, um círculo, ou elipse, tem seu centro posicionado em cada ponto disponível para análise no espaço (mapa) e o seu raio varia de zero até um limite pré-especificado, gerando infinitos círculos. Para cada círculo são assinalados os casos observados e esperados levando em consideração a área fora do círculo. Todos são candidatos a *cluster* e em seguida são avaliados quais são os mais prováveis.

Na varredura espaço-temporal, um cilindro de base circular, ou elíptica, varre todo o espaço. Assim como na varredura espacial, a base é posicionada em cada ponto do espaço e seu raio varia até determinado limite, simultaneamente, a altura do cilindro, que representa o tempo, também varia. Assim, este processo gera infinitos cilindros candidatos a *clusters*.

Para avaliar a verossimilhança dos *clusters* candidatos, podem ser usados alguns modelos de probabilidades, que variam de acordo com a natureza dos dados. Para dados de contagem podem ser usados o modelo de Bernoulli, de Permutação no Espaço-Tempo ou Discreto de Poisson.

Com o modelo Discreto de Poisson, o número de casos em cada localidade segue a distribuição de Poisson. A hipótese nula é a de que o número esperado de casos em cada área é proporcional ao tamanho de sua população, ou das pessoas-ano nessa área (KULLDORFF, 2016).

2.6 AUTOCORRELAÇÃO ESPACIAL GLOBAL E LOCAL

O objetivo do índice de autocorrelação espacial é resumir o grau no qual observações similares, ou dissimilares, tendem a ocorrer próximas umas das outras e pode ser usado para verificar padrões de agrupamento. O índice pode variar em uma direção indicando autocorrelação positiva, na direção oposta indicando autocorrelação negativa, e indicando ausência de autocorrelação quando se aproxima de zero. O índice I de Moran é um índice global de autocorrelação que deriva de uma estatística de produtos cruzados (WALLER; GOTWAY, 2004).

Cada local possui uma relação com seus vizinhos que é estabelecida por uma matriz de proximidade espacial W, que pode ser formada de acordo com vários critérios, sendo o de adjacência o mais utilizado: se um local B fizer fronteira com um local A, então B é vizinho em primeira ordem de A e recebe peso 1 em relação à A, senão peso 0, logo W_{ij}, sendo i=A e j=B é igual a 1. Outros critérios de vizinhança utilizados podem ser o inverso da distância entre os centroides das regiões, ou considerar como vizinhos regiões que estabeleçam um comprimento de fronteira maior que um valor predeterminado (DRUCK, 2004).

Os índices de autocorrelação espacial seguem uma estrutura básica, variando a maneira de computar a similaridade. A equação 1 apresenta a estrutura básica dos índices de autocorrelação espacial global, onde W é a matriz de vizinhança e *sim* é uma medida de similaridade que varia dependendo do índice de autocorrelação espacial.

$$\frac{\sum_{i=1}^{N} \quad \sum_{j=1}^{N} \quad w_{ij}\, sim_{ij}}{\sum_{i=1}^{N} \quad \sum_{j=1}^{N} \quad w_{ij}} \tag{1}$$

Equação 1: Estrutura básica de índices de autocorrelação global (WALLER; GOTWAY, 2004).

2.6.1 Moran Global

O Índice de Moran é muito semelhante ao índice de correlação de Pearson, refletindo uma forma espacialmente ponderada deste coeficiente. Geralmente, o índice é limitado ao intervalo [-1, 1] (mas não restrito a ele), sendo: positivo, se há semelhança entre as taxas em locais próximos; negativo, se há dessemelhança entre as taxas em locais próximos; e $-\frac{1}{N-1}$, se aproximando de zero quanto maior for N (número de locais observados), quando não houver autocorrelação (aleatoriedade espacial) (WALLER; GOTWAY, 2004).

A equação 2 apresenta o cálculo do índice I de Moran, que é derivado da estrutura básica apresentada na fórmula 1. O cálculo consiste na somatória dos produtos entre a matriz de pesos e a matriz de similaridade, dividido pela soma dos pesos, em seguida multiplica-se pelo inverso da variância amostral.

$$I = \frac{1}{s^2} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}\, sim_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}} \quad (2)$$

$$onde\ sim_{ij} = \left(Y_i - \underline{Y}\right)\left(Y_j - \underline{Y}\right), s^2 = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \underline{Y}\right)^2$$

Equação 2: Índice I de Moran (WALLER; GOTWAY, 2004).

A matriz $W_{n\,x\,n}$ é a matriz de pesos da vizinhança. Se i e j são vizinhos, então $w_{i,j} = 1$, se i = j ou i e j não são vizinhos, então $w_{i,j} = 0$. Ressalta-se, como citado anteriormente, que há várias formas de codificar a matriz de pesos e diferentes critérios para considerar um local como vizinho de outro. É comum utilizar o critério de contiguidade Queen, onde A é considerado vizinho de B se compartilharem pelo menos um vértice como fronteira. E a codificação da matriz de pesos é 1, caso compartilhem, e 0, caso contrário.

A similaridade, $sim_{i,j}$, é calculada subtraindo cada valor observado, Y_i e Y_j, pela média do conjunto, $\underline{Y}$. A variância amostral, s^2, é a média da soma dos quadrados das diferenças de cada valor observado, Y_i, pela média do conjunto, $\underline{Y}$. No denominador temos $\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$, representando a soma dos pesos. A seguir, na equação 3, é apresentada a fórmula do I de Moran após substituição e simplificação.

$$I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \left(Y_i - \underline{Y}\right)\left(Y_j - \underline{Y}\right)}{\left(\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}\right) \sum_{i=1}^{N} \left(Y_i - \underline{Y}\right)^2} \quad (3)$$

Equação 3: Índice I de Moran após Simplificação (Fonte: própria).

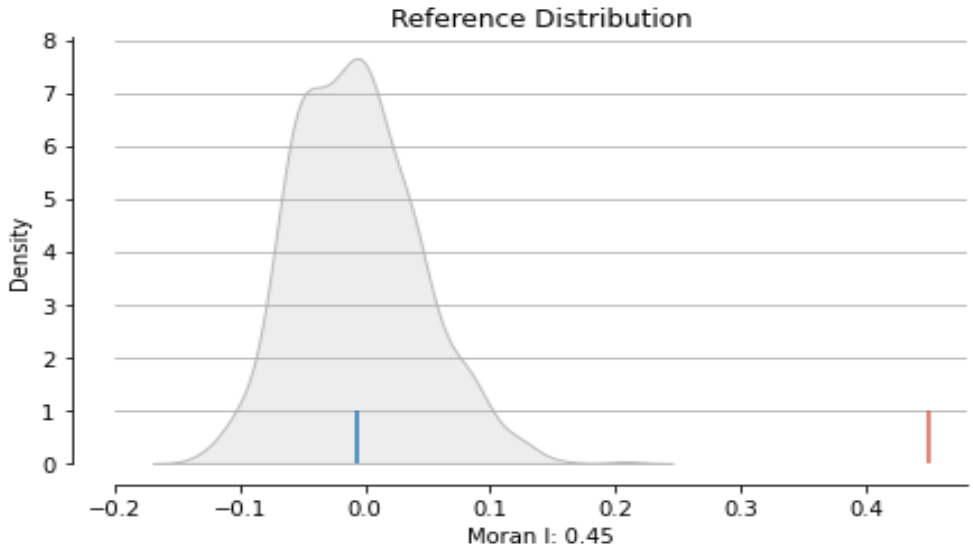
Após o cálculo do Índice, é necessário estabelecer a sua validade estatística. A hipótese nula, H_0, é de aleatoriedade nos dados, e a hipótese alternativa, H_1, é a de que os dados não são aleatórios, apresentando autocorrelação positiva ou negativa. É realizado um teste de pseudo-significância, que consiste em permutar espacialmente os dados analisados (DRUCK, 2004).

O valor esperado do índice de Moran é $EI = -\frac{1}{N-1}$, se aproximando de zero quanto

maior for N. Então, os dados são permutados, por um determinado número de vezes, de maneira a simular aleatoriedade e para cada permutação um I de Moran é calculado. Esses índices geram uma distribuição de referência para a área estudada. Se o valor original de I estiver na região crítica, nas caudas da distribuição de referência, então a hipótese nula de aleatoriedade espacial é rejeitada (WALLER; GOTWAY, 2004) com uma probabilidade de erro equivalente ao *p-value* calculado no teste. O teste é bicaudal e quando a hipótese nula é rejeitada, o I pode estar à direita de EI, indicando padrão de agrupamento, ou à esquerda de EI indicando dissimilaridade nos dados espaciais próximos.

A figura 1 apresenta um exemplo de distribuição de referência após 999 permutações para o I de Moran no ano de 2019 em todo território paraense. O traço azul representa o valor de EI, o I (0,45) está na cauda direita da distribuição (traço vermelho) na região de rejeição da hipótese nula com *p-value* 0,001, indicando padrão de agrupamento nos dados.

Figura 1 – Exemplo de Distribuição de Referência do I de Moran



Fonte: própria.

2.6.2 Moran Local e LISA

O índice global apresenta uma medida da associação espacial para todo um conjunto de áreas, o que é importante para a caracterização inicial da região de estudo. Entretanto, pode ser que nesta região ocorram regimes de associação espacial diferentes e apareçam máximos e mínimos locais, também chamados *hot spots* (pontos quentes) e *cold spots* (pontos frios), e entre estas regiões podem existir outras com baixa significância de agrupamento, que podem

ser consideradas áreas de transição. Estes são padrões de agrupamento importantes que devem ser analisados em maior detalhe. Os índices locais são decomposições dos índices globais e permitem a identificação destes agrupamentos porque produz um índice para cada área (DRUCK, 2004).

A estatística LISA (Indicador Local de Associação Espacial), proposta por Anselin (1995), decompõe o índice global e permite analisar agrupamentos locais que podem ser visualizados através do gráfico de dispersão de Moran I local e do mapa LISA. A equação 4 apresenta a fórmula do Índice de Moran Local.

$$I_i = \left(Y_i - \underline{Y}\right) \sum_{j=1}^{N} w_{ij} \left(Y_j - \underline{Y}\right) \quad (4)$$

Equação 4: Índice I de Moran Local (WALLER; GOTWAY, 2004).

3 METODOLOGIA

Neste capítulo é apresentada a metodologia utilizada na exploração e investigação dos dados de notificações de agravos da LVH. O próprio processo de KDD foi utilizado como metodologia e é detalhado, a seguir, de acordo com a definição dos autores Maimon e Rokach (2010). Devido à organização, clareza, flexibilidade e simplicidade das etapas do KDD, o processo demonstrou ser suficiente e adequado para direcionamento do trabalho. Em seguida, na subseção 3.2, a área de estudo é sucintamente descrita mostrando sua localização geográfica no país. E, por fim, na subseção 3.3, são indicadas as fontes dos dados adquiridos e suas estruturas.

3.1 O PROCESSO DE KDD

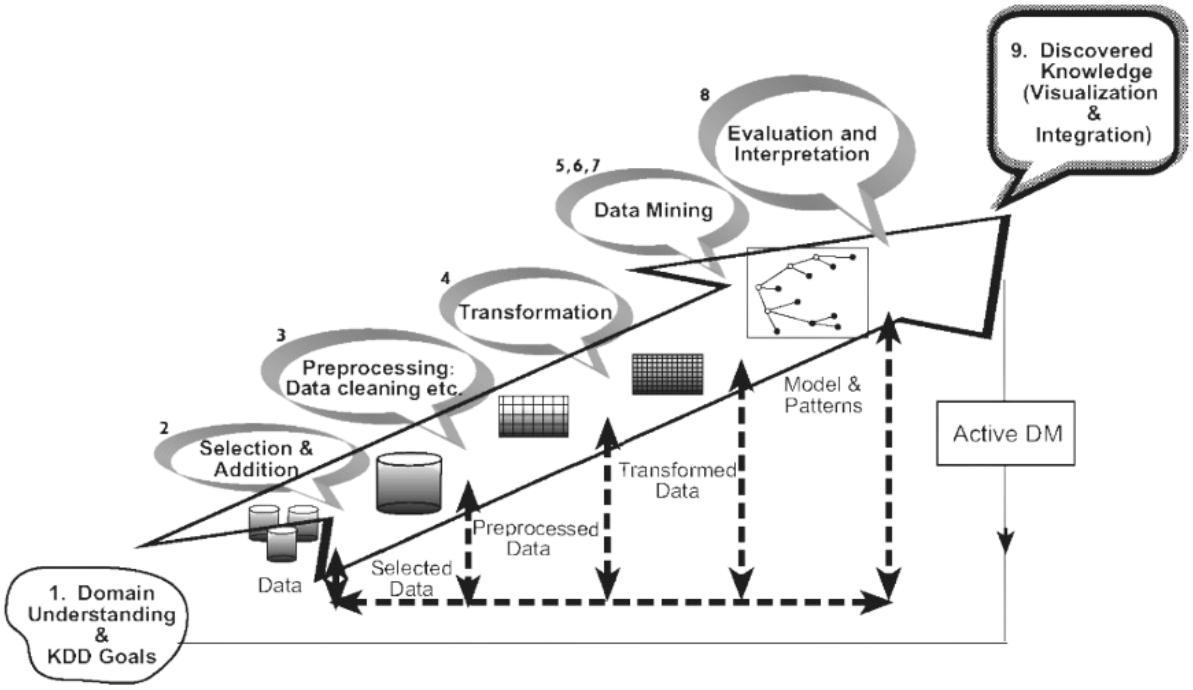
O processo de KDD é organizado, ou seja, existem tarefas bem definidas, entretanto não existe uma receita a ser seguida em cada uma. A forma de lidar com os dados em cada etapa vai depender dos tipos de dados, objetivos, ferramentas e recursos disponíveis.

O processo também é iterativo e interativo. Iterativo porque é possível avançar nas etapas e em seguida retornar para etapas anteriores quantas vezes forem necessárias; o processo pode ser finalizado, produzindo uma parte dos resultados definidos, e em seguida ser reiniciado,

produzindo em cada ciclo um produto (conhecimento). Interativo porque o ser humano interage em todo o processo tomando decisões e avaliando os resultados.

A seguir, são apresentadas as nove etapas do KDD, na visão de Maimon e Rokach (2010), na exposição da figura 2 que resume o processo de KDD.

Figura 2 – O Processo de *Knowledge Discovery* em Bases de Dados



Fonte: Maimon e Rokach (2010).

1) **Compreendendo o domínio da aplicação e definindo os objetivos:** esta etapa consiste em entender e definir os objetivos da aplicação. Algumas decisões sobre transformação e representação dos dados, algoritmos a serem utilizados poderão ser tomadas e executadas nas fases posteriores. Conforme o processo avança talvez esses objetivos sejam ajustados, fazendo com que o fluxo retorne para a fase inicial e que as etapas posteriores também sejam ajustadas. É comum os objetivos não estarem totalmente claros no início, por exemplo, em projetos descritivos. Conforme o processo avança para a fase exploratória e os primeiros *insights* são obtidos, há a possibilidade na qual os objetivos sejam ajustados e o processo retorne para etapas anteriores.
2) **Seleção e adição:** esta é a etapa de seleção e criação de um conjunto de dados a ser explorado. Define-se aqui quais dados são necessários, onde obtê-los, se estão todos disponíveis e se serão necessários dados adicionais. O objetivo é adquirir os dados de

uma ou mais fontes, reuni-los, selecionar os atributos que serão necessários à análise, normalmente ao máximo possível, pois é comum no início do projeto não se saber ainda qual subconjunto de atributos é mais adequado e quais abordagens serão utilizadas. Posteriormente, conforme o processo avança e novos conhecimentos são obtidos, o caráter iterativo do KDD permite retornar a esta fase e refinar a seleção.

3) **Pré-processamento e limpeza dos dados:** nesta etapa são feitos os ajustes para aumentar a confiabilidade dos dados como: limpeza dos dados, tratamento de valores ausentes, remoção de dados discrepantes (*ruídos*). Podem ser utilizados métodos estatísticos ou até mesmo métodos de mineração de dados. Por exemplo, a função "tratar valores ausentes" pode ser implementada através de substituição (ex.: valor ausente por zero), remoção ou até mesmo um algoritmo supervisionado para previsão dos valores ausentes. Desse modo, normalmente, esta é uma etapa que consome bastante tempo.
4) **Transformação:** esta etapa é muito importante porque pode definir se os objetivos do projeto serão alcançados, uma vez que as transformações feitas podem melhorar ou piorar o desempenho dos algoritmos de mineração. Algumas operações que podem ser executadas nesta etapa são a redução de dimensionalidade como a redução de dados vertical (dimensionalidade), redução horizontal (amostragem) e redução de valor (diminuição de valores únicos do atributo). Um exemplo de transformação que pode ser feita é a obtenção de novos atributos através de operações matemáticas e associações entre atributos da base.
5) **Escolha da tarefa de mineração de dados:** de acordo com o objetivo, a tarefa poderá ser do tipo preditiva ou descritiva. Métodos preditivos são utilizados para prever ou rotular, com a maior taxa de acerto possível, através de um modelo treinado com uma amostra dos dados disponíveis para análise. Os métodos descritivos são utilizados para detecção de padrões, agrupamento de dados, busca por regras de associação e visualização dos dados.
6) **Escolha do algoritmo de mineração de dados:** através dos objetivos e dados disponíveis deve ser escolhida a tarefa mais adequada e o melhor método disponível para a tarefa anteriormente escolhida. Suponhamos que a tarefa escolhida seja agrupamento, então considerando as características dos dados a serem analisados e o objetivo a ser alcançado, o algoritmo escolhido deve ser o que traz melhor custo/ benefício. Normalmente, já é sabido quais algoritmos proporcionam melhores resultados em determinado cenário de acordo com trabalhos relacionados e estudos;

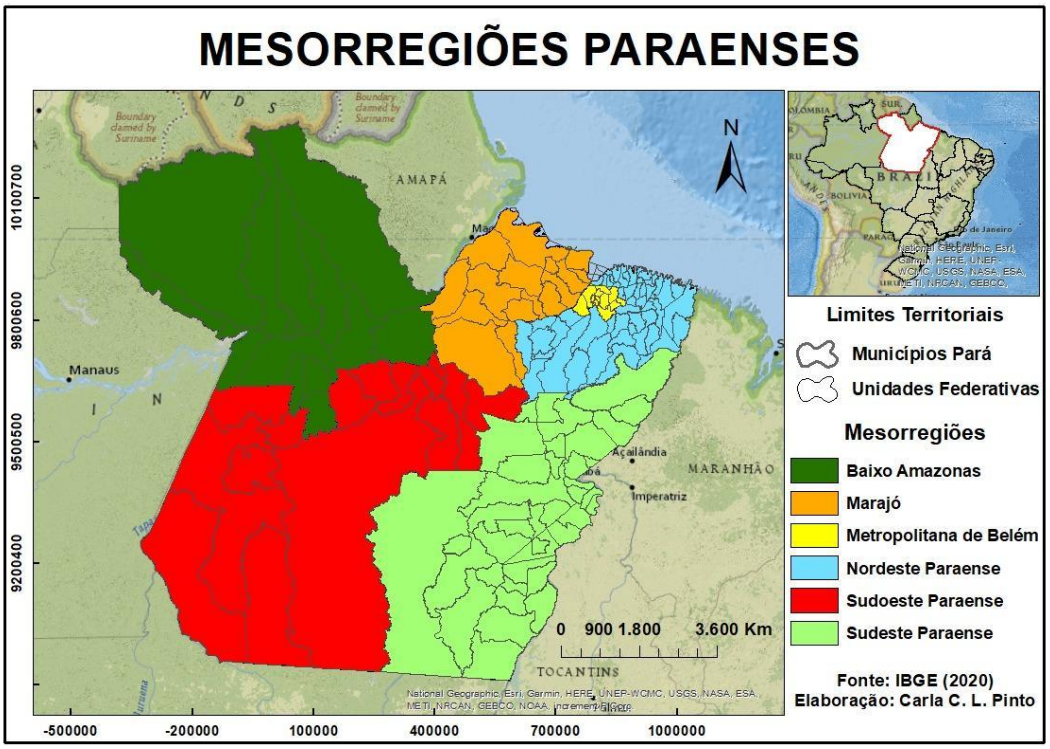
porém, em determinados momentos pode não ser tão simples esta escolha e se faz necessário efetuar testes comparativos, ou seja, avançar e retornar etapas.

7) **Empregando o algoritmo de mineração de dados:** nesta fase podem ocorrer várias execuções do algoritmo onde serão necessários ajustes nos parâmetros até se obter um resultado satisfatório, por exemplo: quantidade de *clusters* em um algoritmo de agrupamento como o *k-means;* tamanho da base de treino e teste em algoritmos preditivos.
8) **Avaliação e interpretação dos resultados:** após a mineração dos dados são verificados se os resultados condizem com os objetivos, se os padrões ou previsões são confiáveis. Pode ser que haja a necessidade de retornar à etapa de transformação ou pré-processamento para fazer ajustes nos dados e executar o fluxo novamente. Também pode ser que os resultados sejam satisfatórios e o projeto avance para a próxima fase.
9) **Usando o conhecimento descoberto: visualização e integração:** o conhecimento adquirido pode ser incorporado a outro sistema para ações futuras. O sucesso desta etapa determina o quão eficaz foi o processo de KDD. O processo se comporta como um *pipeline*: uma vez que o processo foi finalizado e os resultados estão condizentes com os objetivos é possível fazer modificações no sistema e medir os efeitos, realizando experimentos. Exemplo: pode ser que no processo tenham sido utilizados dados apenas de uma região geográfica e seja necessário executar o mesmo processo, mas com a base de dados geograficamente ampliada e comparar os resultados no final. Há vários desafios, pois após a implantação do sistema pode ser que os dados que irão alimentar o processo sejam adquiridos em tempo real e que tenham tipos de dados diferentes do previsto, atributos novos, com valores fora do domínio ou faltantes.

3.2 ÁREA DE ESTUDO

O estado do Pará fica na região Norte do Brasil, sendo o segundo maior estado do país em extensão territorial, somente menor que o Amazonas, com área de 1.247.955,238 km^2 (IBGE, 2017). Faz limites ao norte com o Amapá, ao noroeste com Suriname, Guiana e Roraima, ao oeste com Amazonas, ao sul com Mato Grosso, ao sudeste com Tocantins, leste com Maranhão e nordeste com Oceano Atlântico. Possui população e densidade demográfica estimadas, para o ano de 2021, de 8.777.124 pessoas e 7,03 habitantes por km². O estado possui 144 municípios divididos em seis mesorregiões, conforme a imagem a seguir.

Figura 3 – Mesorregiões Paraenses



Fonte: Dados IBGE (2020).

3.3 INDICADORES EPIDEMIOLÓGICOS

Os indicadores epidemiológicos são importantes para caracterizar uma situação de saúde em uma região. O Caderno de Indicadores das Leishmanioses Tegumentar e Visceral do Ministério da Saúde (BRASIL *et al.*, 2018) demonstra como calcular 13 indicadores para a LVH, que são:

1. Total de casos;
2. Taxa geral de incidência;
3. Proporção de casos confirmados por critério laboratorial;
4. Proporção de casos na faixa etária menor de 5 anos;
5. Proporção de casos na faixa etária de 50 anos ou mais;
6. Proporção de casos em coinfectados por HIV;
7. Proporção de casos que evoluíram para cura clínica;
8. Número de óbitos por LV;
9. Taxa de letalidade por LV;
10. Taxa de letalidade por LV em coinfectados por HIV;

11. Taxa de letalidade por LV na faixa etária menor de 5 anos;
12. Taxa de letalidade por LV na faixa etária de 50 anos ou mais;
13. Proporção de casos de LV com evolução ignorada ou em branco.

Os indicadores que foram utilizados na análise dos dados foram: i) total de casos e ii) taxa geral de incidência.

3.3.1 Taxa Geral de Incidência

A taxa geral de incidência – *t* – para um determinado local é o total de casos registrados no ano de notificação dividido pela estimativa populacional do mesmo ano e multiplicado por 100.000. Ou seja, é a quantidade de casos confirmados para cada 100.000 habitantes. A limitação deste indicador, segundo o Caderno de Indicadores, é a inclusão de casos alóctones no numerador do indicador: uma pessoa pode ter se infectado em um local A, mas residir em um local B, entretanto o denominador do indicador contempla apenas a estimativa populacional do local A. A equação 5 apresenta a fórmula do cálculo do indicador.

$$t = \frac{Total\ de\ Casos\ por\ Local\ de\ Infecção\ no\ Ano\ de\ Notificação}{População\ Total\ do\ Local\ no\ Ano\ de\ Notificação}\ x\ 100.000 \quad (5)$$

Equação 5: Fórmula do Cálculo da Taxa de Incidência (BRASIL *et al*., 2018).

3.3.2 Total de Casos

É o número total de casos novos confirmados de LVH por local provável de infecção (UF, município, região administrativa ou localidade) no ano de notificação.

3.4 FONTE E DESCRIÇÃO DOS DADOS

As notificações de casos de LVH são geridos pelo SINAN (Sistema de Informação de Agravos de Notificação) e estão disponíveis no site do DataSUS.[10] Os dados de estimativas populacionais, necessários para o cálculo do indicador epidemiológico, também estão disponíveis no site do DataSUS. A fonte é "Base Populacional - IBGE" e o arquivo utilizado

[10] Transferência de Arquivos - Datasus. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/. Acesso em: 27 jan. 2022.

foi “Estimativas TCU - 1992 até 2019”.

Os dados sobre os municípios (códigos, nomes, mesorregiões, microrregiões, polígonos e coordenadas) foram obtidos no site do IBGE e integrados à base de notificações com o objetivo de facilitar a análise e visualização dos dados. Em “IBGE: Divisão Territorial”,[11] na planilha RELATORIO_DTB_BRASIL_MUNICIPIO.xls, estão presentes informações gerais sobre os municípios. Assim, também, estão disponíveis os dados de polígonos de áreas dos municípios brasileiros estão disponíveis em “IBGE: Malhas Territoriais”.[12]

Os atributos presentes na base de dados de notificações são oriundos das fichas de investigação de LVH, mantidas pelos profissionais das Unidades Básicas de Saúde. Nesta ficha existem 55 campos e espaço para observações divididos em 9 seções que são:

1. Dados Gerais.
2. Notificação Individual.
3. Dados de Residência.
4. Antecedentes Epidemiológicos.
5. Dados Clínicos.
6. Dados Laboratoriais/Classificação do Caso.
7. Tratamento.
8. Conclusão.
9. Informações Complementares e Observações.

Estas seções capturam informações sobre os locais de notificação, residência, infecção, algumas informações relacionadas ao diagnóstico e dados clínicos, como sintomas. Também informações sobre o tratamento e evolução do caso, data de início de tratamento e de primeiros sintomas, droga administrada, se evoluiu para cura, se é caso novo, ou recidiva, e observações gerais.

Alguns campos são essenciais para a identificação de quais casos serão, ou não, utilizados no cálculo de indicadores:

- CLASSI_FIN: indica se o caso foi confirmado (1), ou descartado (2). Se o caso tiver

[11] IBGE: Divisão Territorial. Disponível em: https://geoftp.ibge.gov.br/organizacao_do_territorio/estrutura_territorial/divisao_territorial/2018/DTB_2018.zip. Acesso em: 27 jan. 2022.

[12] IBGE: Malhas Territoriais. Disponível em: https://geoftp.ibge.gov.br/organizacao_do_territorio/malhas_territoriais/malhas_municipais/municipio_2020/Brasil/BR/BR_Municipios_2020.zip. Acesso em: 27 jan. 2022.

diagnóstico parasitológico, imunológico ou outro, então automaticamente é um caso confirmado. Se todos os tipos de diagnósticos deram negativo e não houver nenhuma manifestação clínica, automaticamente o caso é descartado. Observou-se que alguns casos não respeitam esta regra, entretanto nenhuma correção foi feita na base. Assim, foram selecionados apenas os casos com CLASSI_FIN igual a 1.

- NDUPLIC_N: indica se o caso é duplicado, ou não. Quando 0, a duplicidade não foi identificada, quando 1 não é duplicidade, se 2 então é duplicidade. Todos os casos com NDUPLIC_N igual a 2 foram excluídos.
- ENTRADA: indica se o caso é novo (1), recidiva (2), transferência de outro município ou estado (3), ou possui valor ignorado (9). Um caso é considerado recidiva quando o paciente apresentou manifestações clínicas em um período de até 12 meses após a cura clínica da LVH (BRASIL, 2019). Foram considerados apenas os casos novos, com valor igual a 1.
- CO_MN_INF: indica o código IBGE do município provável de infecção. Os campos que indicam a localidade (notificação, residência e infecção) podem estar nulos ou possuir valor ignorado, que é quando os municípios são iniciados com o código do estado e terminados com 0000.

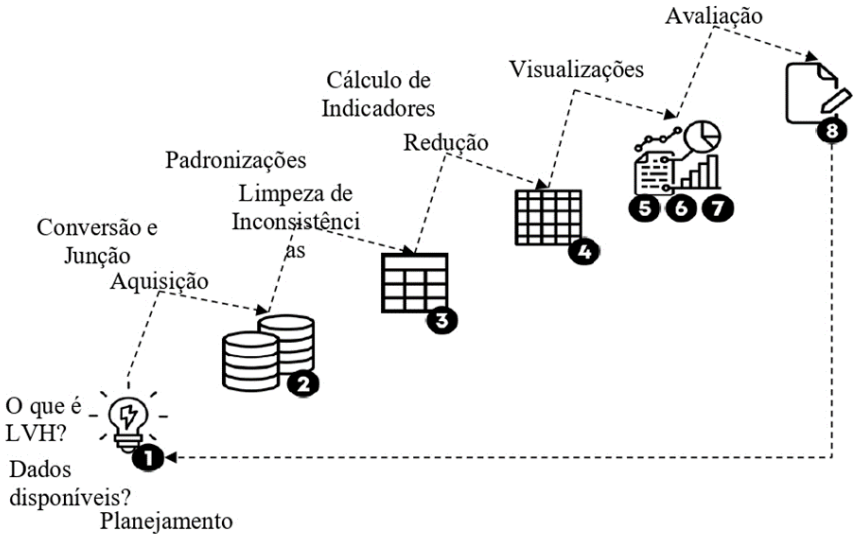
4 IMPLEMENTAÇÃO

A implementação deste trabalho foi executada em três iterações que são apresentadas neste capítulo. Cada subcapítulo descreve em detalhes as operações que foram realizadas nos dados em cada iteração. Os resultados das iterações são compilados e apresentados no Capítulo V, Testes e Avaliações.

4.1 PRIMEIRA ITERAÇÃO

A primeira iteração consistiu em pesquisas para entender a problemática da LVH, identificação prévia de dados disponíveis e planejamento para aquisição e exploração dos dados. A figura 4 resume o processo.

Figura 4 – Fluxograma das Etapas da Primeira Iteração



Fonte: própria

4.1.1 Etapa I: Entendimento do Negócio

Embora não se soubesse a princípio a dimensão do problema da LVH no estado do Pará, foi constatado através de pesquisas na *web* que a LVH é um problema de saúde pública

relevante em várias regiões do Brasil, além do Pará. Em seguida foi verificada e confirmada a disponibilidade de dados.

O próximo passo foi obter conhecimento geral sobre a doença, portanto foram consultados materiais como notícias, estudos entomológicos, artigos e manuais. Alguns exemplos de consulta são o artigo de Silveira *et al.* (2016), mostrando a evolução da doença no território brasileiro e apresentando informações sobre o seu ciclo e manuais como o Guia de Vigilância em Saúde da Leishmaniose Visceral (BRASIL *et al.*, 2016) e o Caderno de Indicadores das Leishmanioses (BRASIL *et al.*, 2018), que explica e demonstra como calcular os principais indicadores da doença.

Nesse âmbito, esta etapa foi essencial para o conhecimento geral do problema e possibilitou melhor entendimento das variáveis presentes nas bases.

4.1.2 Etapa II: Seleção e Adição de Dados

4.1.2.1 Aquisição

Nesta etapa a verificação de disponibilidade e fonte dos dados foi ampliada, identificando-se onde e como obter os dados. As bases de notificações e de estimativas populacionais foram adquiridas no portal SINAN e consistem em arquivos anuais comprimidos nos formatos DBC e DBF. O formato DBC do DataSUS consiste em uma compactação do formato DBF, e foi criado pelo Ministério da Saúde (PETRUZALEK, 2016).

Os dados do IBGE também foram adquiridos e consistem em planilhas no formato XLS (formato da *Microsoft*) e arquivos de polígonos no formato SHP[13] (*Shapefile*). Este, é um formato de arquivo aberto, que representa informações geoespaciais através de vetores e permite a construção de mapas e utilização dos dados em diferentes GIS, Sistemas de Informação Geoespacial (*Geospatial Information Systems)*.

[13] ESRI Shapefile Technical Description. Disponível em https://www.esri.com/content/dam/esrisites/sitecore-archive/Files/Pdfs/library/whitepapers/pdfs/shapefile.pdf. Acesso em: 8 mar. 2022.

4.1.2.2 Conversão

Há ferramentas de tabulação para os dados nos formatos DBC e DBF, que são o TabNet,[14] tabulador online, e o TabWin,[15] programa tabulador que funciona no sistema operacional Windows, e que podem ser utilizados na exploração inicial e no cálculo de indicadores. Entretanto, para maior controle de quais notificações incluir, ou não, na análise, e devido à necessidade de flexibilidade no tratamento, integração e mineração dos dados durante o processo de KDD, optou-se por converter os dados para o formato CSV[16] (*Comma Separated Values Format*), que é um formato aberto e muito utilizado por funcionar em vários programas tabuladores.

Para a conversão dos dados em DBC para CSV, foi utilizada uma biblioteca em R, *readdbc* (PETRUZALEK, 2016), que foi desenvolvida especialmente para converter este tipo de dado do DataSUS. Os arquivos em DBF foram convertidos através de uma biblioteca *Python, dbfread.*[17]

Após a conversão, os dados anuais foram reunidos nas bases de estimativa populacional e de notificações, representando todo o período de 2007 a 2019.

4.1.3 Etapa III: Pré-Processamento e Limpeza dos Dados

Na fase de limpeza, a principal operação realizada na base de notificações foi a de limpeza de inconsistências por correção de erros.

4.1.3.1 Inconsistências de Locais

Existem três tipos de localização na base: local de residência, local de notificação e local provável de infecção. Cada um tem um campo para código de município e de UF. É uma inconsistência um caso com município provável de infecção ter os dois primeiros dígitos indicando ser de um determinado estado e o código de UF de infecção ser de outro estado. Ex.:

[14] Informações de Saúde (TABNET) - DataSUS. Disponível em: https://datasus.saude.gov.br/informacoes-de-saude-tabnet/. Acesso em: 8 mar. 2022.
[15] Tabwin - SINANWEB. 7 mar. 2016. Disponível em: http://portalsinan.saude.gov.br/sistemas-auxiliares/tabwin. Acesso em: 8 mar. 2022.
[16] RFC 4180 - Common Format and MIME Type for Comma-Separated. Disponível em: https://tools.ietf.org/html/rfc4180. Acesso em: 8 mar. 2022.
[17] dbfread - Read DBF Files with Python — dbfread 2.0.7 documentation. Disponível em: https://dbfread.readthedocs.io/. Acesso em: 21 fev. 2022.

CO_MN_INF = 150002 e CO_UF_INF = 17. Neste caso vale o que o código do município de infecção indica. Sendo assim, esta não é uma correção crítica, pois não influencia no cálculo do indicador, visto que os indicadores utilizam o código de município de infecção. Entretanto, foi realizada para facilitar a manipulação dos dados.

4.1.3.2 Atributos de datas

Algumas datas estão em formato inadequado, por exemplo, o ano sendo representado pelos dois últimos dígitos somente. Desta forma, não é possível realizar cálculos com as datas que possam revelar o tempo entre primeiros sintomas e início do tratamento, idade do paciente no dia da notificação, dentre outros. Não é uma correção crítica para esta iteração, visto que a princípio serão utilizados apenas os indicadores.

4.1.3.3 Códigos de municípios

As bases do SINAN utilizam o código do IBGE com seis dígitos, atualmente o código possui sete dígitos. Os dois primeiros são referentes ao estado e o último é um dígito verificador. Os seis dígitos são suficientes para identificar e é mais simples descartar o último dígito dos registros que possuem sete dígitos do que corrigir todos os códigos que possuem seis dígitos. Esta operação consistiu na padronização do código IBGE nas bases de notificações e população para que os dados de notificações individuais fossem cruzados corretamente.

4.1.3.4 Codificação

Os dados foram convertidos para o tipo adequado, por exemplo, datas de formato *string* para *datetime*, variáveis quantitativas discretas para o tipo *int* (inteiro) e contínuas para *float*, não sendo uma operação crítica.

4.1.4 Etapa IV: Transformação

Nesta etapa foram realizadas operações de redução de dados e criação de novos atributos, os indicadores.

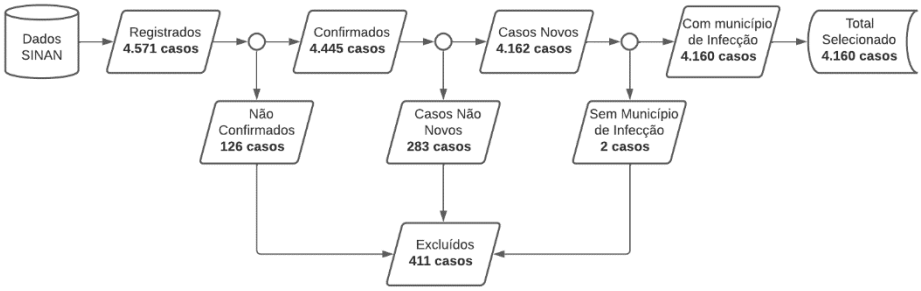
4.1.4.1 Redução dos dados

Para os cálculos dos indicadores – taxa geral de incidência e total de casos – é necessário selecionar apenas os casos confirmados (CLASSI_FIN = 1) e novos (ENTRADA = 1), que possuem local de infecção, código do IBGE, válido e diferente de ignorado. Esta é a operação de redução horizontal direta, quando se selecionam amostras dos dados baseado em algum critério não aleatório, neste caso o valor de algumas variáveis.

A base de notificações possui 4.571 casos registrados para o estado do Pará no período de 2007 a 2019, dos quais nenhum é duplicado. Após a seleção apenas dos casos confirmados, CLASSI_FIN igual a 1, restaram 4.445 amostras (126 amostras foram excluídas). Em seguida foram selecionados dentre os confirmados, apenas os casos novos (ENTRADA igual a 1), totalizando 4.162 casos, sendo 283 casos excluídos.

Dentre os casos novos e confirmados, foram encontrados 2 casos sem código de município de infecção e estes foram removidos. Nenhum município ignorado foi encontrado. Logo, a amostra analisada foi de 4.160 casos novos, 93% de todos os casos confirmados no estado do Pará entre 2007 e 2019. Também foram removidas nove variáveis (colunas) desnecessárias, redução vertical dos dados. A figura 5 resume o processo de seleção dos dados de notificações.

Figura 5 – Fluxograma da Seleção dos Dados



Fonte: própria.

4.1.4.2 Cálculo dos Indicadores

Foram calculados dois indicadores: taxa geral de incidência anual e total de casos anuais, conforme explicitado no Capítulo III, item 3.3 Indicadores Epidemiológicos.

4.1.5 Etapa V: Escolha da Tarefa de Mineração

Por se tratar da primeira iteração, ainda não havia sido feita uma exploração inicial em busca de padrões nos dados, apenas em relação às inconsistências da base. De acordo com os questionamentos que surgiram na fase de projeto, é um objetivo entender se o total de casos por 100.000 habitantes esteve distribuído uniformemente pelo território no período ou se houve concentração em alguns locais. Logo, a tarefa escolhida foi descritiva.

4.1.6 Etapa VI: Escolha do Algoritmo de Mineração de Dados

Nesta etapa optou-se em realizar a primeira extração de conhecimento das bases através de técnicas exploratórias como manipulação, segmentação de dados e visualização através de gráficos, para a identificação de padrões na distribuição de casos pelo território.

4.1.7 Etapa VII: Empregando o Algoritmo de Mineração de Dados

O estado foi dividido em subgrupos (mesorregiões, microrregiões e municípios) e suas séries históricas de taxas de incidência foram comparadas entre si com o objetivo de verificar se havia correlação entre estes subgrupos através de gráficos de linha e correlogramas.

Foi constatada, visual e estatisticamente, a correlação entre as taxas de incidência entre diferentes locais através de gráficos de linhas, comparando a evolução das taxas anuais, e de medidas de correlação Pearson.

4.1.8 Etapa VIII: Avaliação e Interpretação dos Resultados

Esta exploração permitiu identificar que os casos de LVH por 100.000 habitantes, no período de estudo, não estavam proporcionalmente distribuídos no território; que no início do período analisado os casos se concentraram no Nordeste paraense, Região Metropolitana de

Belém e Marajó; que no final do período os casos se concentraram no Sudeste Paraense e que nesta última região alguns municípios tiveram altas exorbitantes nas taxas.

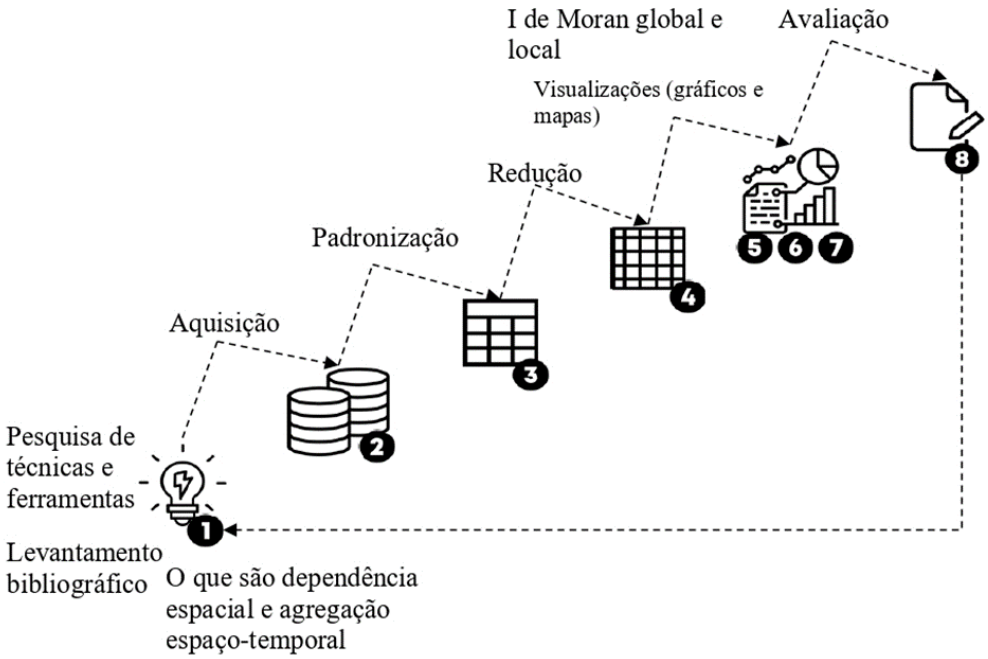
Entretanto, este não é o melhor método para verificar a correlação em dados espaciais. É necessário considerar o fenômeno de dependência espacial. Também existem outras formas mais adequadas para visualizar o fenômeno de agregação espaço-temporal, quando há excessos de casos simultaneamente nas dimensões espacial e temporal.

Portanto, este processo retornou para a fase de entendimento, onde novas pesquisas foram feitas em relação ao conceito de agregação espaço-temporal e dependência espacial, e os objetivos e entendimento do negócio (projeto) foram atualizados e executados na segunda iteração.

4.2 SEGUNDA ITERAÇÃO

A segunda iteração consistiu em novas pesquisas, levantamento bibliográfico para identificação de trabalhos semelhantes e pesquisa das técnicas identificadas em trabalhos relacionados. Assim, a figura 6 resume o processo.

Figura 6 – Fluxograma das Etapas da Segunda Iteração



Fonte: própria

4.2.1 Etapa I: Entendimento do Negócio

Novas pesquisas foram feitas com o objetivo de compreender melhor o conceito de agregação de dados no espaço e tempo. Assim, foram lidos alguns trabalhos semelhantes encontrados em repositórios de trabalhos científicos, como o Periódico Capes, e o entendimento e objetivos foram atualizados.

Baseado no artigo de Werneck e Struchiner (1997)**,** que mostra técnicas de visualizações de tendência de agrupamentos e faz uma revisão sobre alguns métodos estatísticos para detecção de aglomerados espaciais, o objetivo foi continuar a exploração visualizando a agregação na dimensão temporal, espacial e espaço-temporal. Além disso, estimar através do Índice Global de Moran a tendência de agrupamentos na região e visualizar a decomposição do índice através da Estatística LISA.

Para tanto, foram pesquisadas novas ferramentas para visualização de dados espaciais e estatística espacial. Entre as opções encontradas optou-se pelas bibliotecas, *pysal* (estatística) e *geopandas* (visualização). Também foi necessário verificar a disponibilidade de dados adicionais como os de polígonos, ou malhas territoriais, que estão disponíveis no site do IBGE.

4.2.2 Etapa II: Seleção e Adição de Dados

Os dados adicionais levantados na fase anterior, cuja descrição e fonte se encontram no Capítulo III, item 3.4, Fonte e Descrição dos Dados, foram obtidos.

4.2.3 Etapa III: Pré-Processamento e Limpeza dos Dados

Os dados sofreram alterações mínimas, apenas padronização do código do município para ficarem compatíveis com as outras bases.

4.2.4 Etapa IV: Transformação

A operação de redução horizontal direta foi realizada, selecionando apenas os polígonos do estado do Pará para as fases seguintes.

4.2.5 Etapa V: Escolha da Tarefa de Mineração

De acordo com os produtos da Etapa I desta iteração, ficou definido que seriam aplicadas técnicas de visualização de tendência de agrupamentos, verificando se há excesso na dimensão temporal, espacial e em ambas simultaneamente. Também ficou estabelecida a utilização de métodos estatísticos para quantificar e visualizar a agregação de casos de LVH no território paraense. Logo, as tarefas são descritivas.

4.2.6 Etapa VI: Escolha do Algoritmo de Mineração de Dados

A exploração foi realizada em um *Jupyter Notebook*[18] através das bibliotecas *matplotlib,*[19] *libpysal,*[20] *splot*[21] e *esda*[22] para cálculo do I de Moran Global e Local, visualização dos *clusters* e construção de gráficos.

4.2.7 Etapa VII: Empregando o Algoritmo de Mineração de Dados

De acordo com Werneck e Struchiner (1997, p. 612):

> A agregação espaço-temporal pode ser entendida como uma forma de não-aleatoriedade na distribuição da doença: em que, entre eventos próximos no tempo, existe um excesso não esperado de eventos que estão também próximos no espaço (McAullife & Afifi, 1984). Este conceito é distinto do de agregação espacial e temporal, e também tem sido denominado interação espaço-temporal (Knox, 1991; Jacquez et al., 1996). De fato, agregação espaço-temporal pode ocorrer na ausência de agregação espacial e temporal, ou mesmo estar ausente quando existe agregação nas duas dimensões (Estève et al., 1994).

Seguindo o exemplo da publicação, foram criadas algumas visualizações para identificar a agregação espacial, temporal e espaço-temporal.

A agregação espacial pode ser visualmente identificada plotando o total de casos, ou taxa, de cada município em determinado período (ano, triênio, média de todo o período). Uma forma de se fazer isso é através de mapas coropléticos. Foram plotados um mapa com a média

[18] Jupyter Notebook. Disponível em: https://jupyter.org/. Acessado em 21 fev. 2022.
[19] Matplotlib — Visualization with Python. Disponível em: https://matplotlib.org/. Acesso em: 21 fev. 2022.
[20] Python Spatial Analysis Library Core — libpysal v4.6.0 Manual. Disponível em: https://pysal.org/libpysal/. Acesso em: 21 fev. 2022.
[21] splot - PyPI." https://pypi.org/project/splot/. Acesso em: 21 fev. 2022.
[22] Exploratory Spatial Data Analysis — esda v2.4.1 Manual - PySAL. Disponível em: https://pysal.org/esda/. Acesso em: 21 fev. 2022.

das taxas de incidência dos treze anos (2007-2019) e outro com a média do último triênio (2017-2019), as taxas de incidência anuais dos municípios foram estratificadas em quatro intervalos iguais, ou seja, o intervalo entre a menor e a maior taxa foi dividido em quatro intervalos iguais e cada intervalo é representado por uma cor. Dessa forma, fica fácil observar no mapa os municípios que estão em cada intervalo.

A agregação temporal pode ser observada plotando o total de casos, ou taxa, de cada ano em determinado espaço. Foi utilizado um gráfico em barras com evolução da taxa de incidência média durante o período de treze anos para todo o estado.

Para visualizar a agregação nas duas dimensões utilizou-se o gráfico de barras empilhadas para mostrar a taxa de incidência média anual do estado segmentada pelas mesorregiões paraenses.

O I de Moran global foi positivo em todos os anos, sendo maior em 2019 (0,45). A decomposição do índice global foi visualizada através do mapa LISA e grupos de municípios com taxas altas próximos de outros com taxas altas (alta-alta) foram localizados no nordeste e sudeste do estado.

4.2.8 Etapa VIII: Avaliação e Interpretação dos Resultados

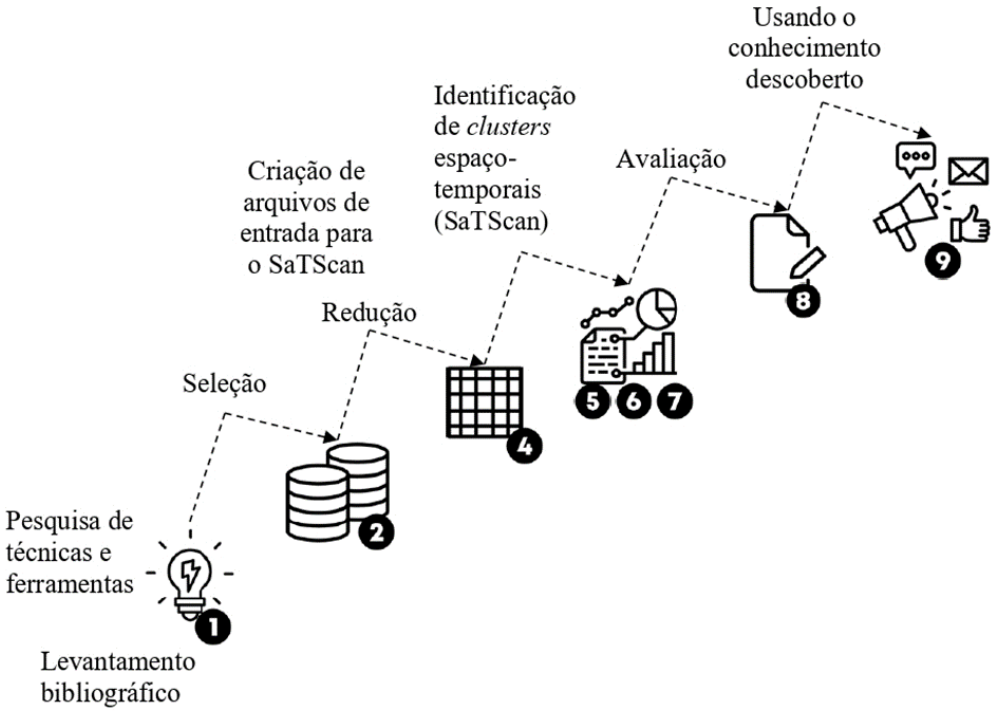
Através da exploração anterior foi verificado que existe tendência de agrupamento no território, indicado pelo índice de Moran positivo para todo o período. Em relação ao último triênio, 2017-2019, o mapa LISA mostrou que o sudeste paraense foi a única região que apresentou municípios autocorrelacionados com taxas alta-alta.

Neste ponto foi constatado que seria interessante utilizar um método que não apenas detectasse os locais com excessos de casos em um período pré-determinado, mas que automaticamente procurasse os agrupamentos e períodos em que ocorreram. Portanto, foi necessária uma nova iteração, voltando para a primeira etapa, onde foram feitas novas pesquisas e os objetivos foram novamente alinhados.

4.3 TERCEIRA ITERAÇÃO

A terceira iteração consistiu principalmente em entender e aplicar a Estatística de Varredura Espaço-Temporal através do software SaTScan. A figura 7 resume o processo.

Figura 7 – Fluxograma das Etapas da Terceira Iteração



Fonte: própria

4.3.1 Etapa I: Entendimento do Negócio

Foram realizadas novas pesquisas e encontrados artigos que tratavam da agregação espaço-temporal da LVH. A varredura espaço-temporal se mostrou adequada para o problema e tipo de dado disponível. Este método percorre todo o espaço através de um cilindro móvel. A base, que é flexível, seleciona pontos no espaço, a altura do cilindro, também flexível, seleciona o tempo. A estatística de varredura seleciona áreas, através do cilindro, e compara estas seleções com a área externa ao cilindro. Dessa forma, a varredura consegue encontrar áreas com excesso de ocorrências, proporcionais ao total de habitantes, e também em quais períodos o agrupamento ocorreu.

O objetivo desta iteração é, através da estatística de varredura espaço-temporal implementada no *software* SaTScan, realizar uma varredura nas taxas de incidência do estado do Pará para todo o período e identificar quais áreas e em quais períodos ocorreram agrupamentos espaço-temporais da LVH.

4.3.2 Etapa II: Seleção e Adição de Dados

Nesta iteração não foi necessário adquirir novos dados, apenas selecionar aqueles já adquiridos. Juntamente com os arquivos de malhas territoriais, disponibilizados pelo IBGE, há um arquivo de localidades (relação de todas as localidades do Brasil), onde estão alguns atributos necessários que são as coordenadas das cidades paraenses.

O algoritmo de varredura espaço-temporal (SaTScan) precisa de um ponto de início para posicionar a base do cilindro. Por padrão ele considera o centróide do município, entretanto pode ser que o centro não seja a região mais densa populacionalmente. Existe a opção de indicar através de um arquivo de grade quais são os pontos que o algoritmo deve considerar para iniciar a varredura.

As coordenadas das cidades dos municípios foram utilizadas como arquivo de grade na análise com o SaTScan. Nesta etapa, a base com os atributos de coordenadas foi selecionada.

4.3.3 Etapa III: Pré-Processamento e Limpeza dos Dados

Não foi necessário pré-processamento.

4.3.4 Etapa IV: Transformação

No arquivo de grade foi feita uma redução vertical, extraindo apenas os códigos dos municípios, latitude e longitude, e horizontal, apenas em cidades paraenses.

Foram criados conjuntos novos de população e de casos, conforme o formato exigido pelo algoritmo, a partir dos arquivos de população e total de casos já processados na primeira iteração.

4.3.5 Etapa V: Escolha da Tarefa de Mineração

Como já adiantado no início desta iteração, a tarefa é descritiva e busca por agrupamento nos dados.

4.3.6 Etapa VI: Escolha do Algoritmo de Mineração de Dados

O algoritmo é o de varredura espaço-temporal, modelo de probabilidade Discreto de Poisson, implementado pelo *software* SaTScan.

4.3.7 Etapa VII: Empregando o Algoritmo de Mineração de Dados

A análise espaço-temporal realizada pelo SaTScan pode ser retrospectiva, quando analisa dados históricos, ou prospectiva, quando a análise é feita periodicamente conforme novos dados são adquiridos (KULLDORFF, 2016). O tipo de análise realizada foi espaço-temporal retrospectiva. O modelo de probabilidade foi o Poisson Discreto. A análise buscou por áreas com taxas altas e utilizou o ano como unidade de agregação temporal com comprimento de três anos. A análise retornou apenas *clusters* circulares com até 50% da população de risco, esta é uma opção padrão do *software*.

4.3.8 Etapa VIII: Avaliação e Interpretação dos Resultados

O algoritmo de varredura encontrou três *clusters*: o primeiro entre 2014 e 2019, na mesorregião do sudeste paraense; o segundo entre 2007 e 2013 nas mesorregiões nordeste paraense e Metropolitana de Belém; e o terceiro na mesorregião do Marajó entre 2007 e 2013.

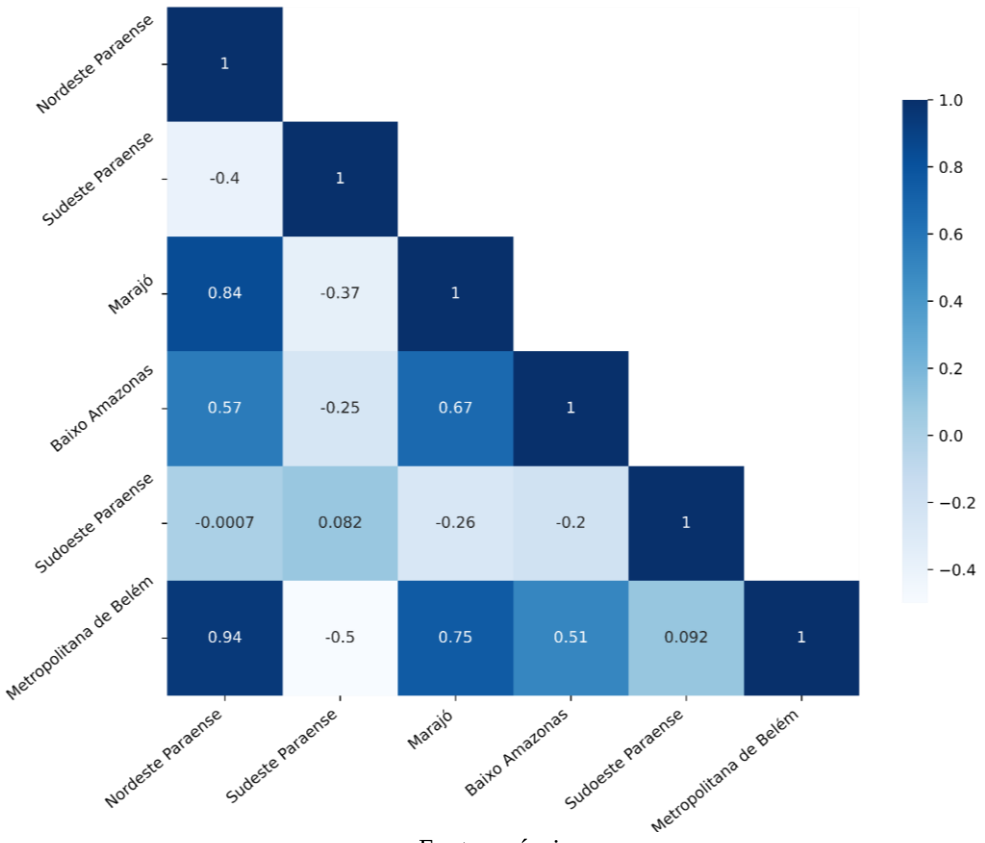
Observa-se que os agrupamentos identificados pela estatística de varredura são similares aos mapeados pelo Índice de Moran Local. O SaTScan complementou a análise, pois, além do período em que os agrupamentos ocorreram, apresentou informações extras como raio do *cluster*, número de municípios, risco relativo, tamanho da população e relação entre casos observados e esperados.

Esta análise, até esta iteração, responde às perguntas iniciais da pesquisa. O Capítulo V compila os resultados do KDD com auxílio de gráficos e tabelas. O Capítulo VI conclui e discute os resultados.

5 TESTES E AVALIAÇÕES

A correlação Pearson entre as séries temporais das mesorregiões foi calculada e resultou no correlograma (Figura 8). A Região Metropolitana de Belém, o Nordeste Paraense e o Marajó, apresentaram forte correlação positiva entre si. O Sudeste apresentou correlação positiva muito fraca com o Sudoeste, 0,08, e negativa, entre -0,5 e -0,25, com as outras mesorregiões. O Baixo Amazonas apresentou correlação positiva média, entre 0,51 e 0,67, com o Nordeste, Marajó e Metropolitana de Belém.

Figura 8 – Correlação Pearson entre as Séries Temporais das Mesorregiões Paraenses



Fonte: própria

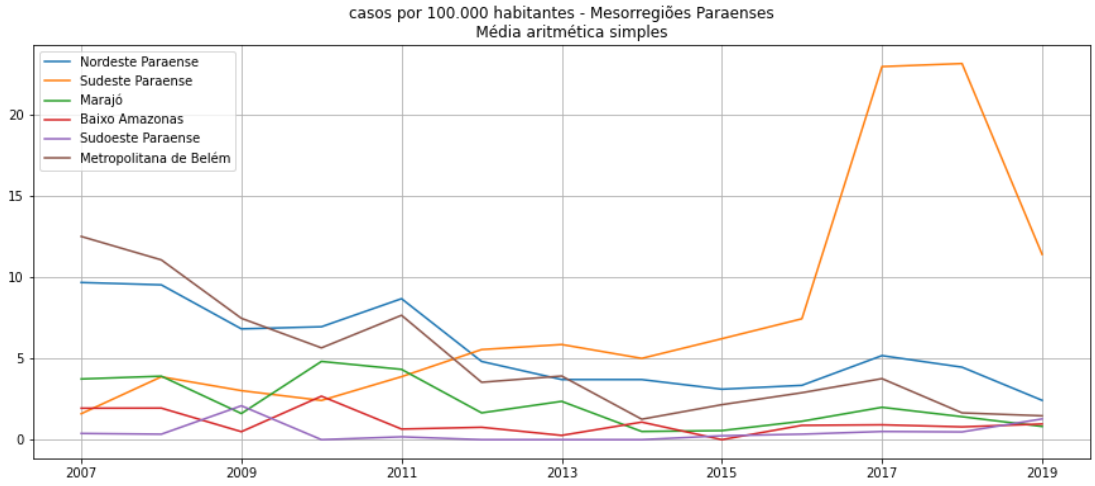
Em suma, as mesorregiões do Baixo Amazonas, Metropolitana de Belém, Marajó e Nordeste paraense apresentaram correlação positiva entre si. Sudeste e Sudoeste do Pará não apresentaram correlação significativa, positiva ou negativa, entre as demais.

No início do período de estudo as mesorregiões Metropolitana de Belém e Nordeste paraense (linhas marrom e azul), na figura 9, lideraram com as maiores taxas. A partir de 2011,

as taxas da mesorregião Sudeste paraense (linha amarela) aumentaram e as taxas da Metropolitana e Nordeste do Pará começaram a diminuir, fazendo do Sudeste do Pará protagonista em relação à LVH. Nos últimos quatro anos do período analisado, o Sudeste do estado do Pará enfrentou aumento abrupto nas taxas, de 2016 para 2017 foi um aumento de aproximadamente 300%. O ápice foi no ano de 2018.

O gráfico de linhas (Figura 9), indicou que os casos de LVH não estavam distribuídos de maneira uniforme no estado, especialmente se tratando dos últimos anos. E a região Sudeste deve ser analisada em mais detalhes.

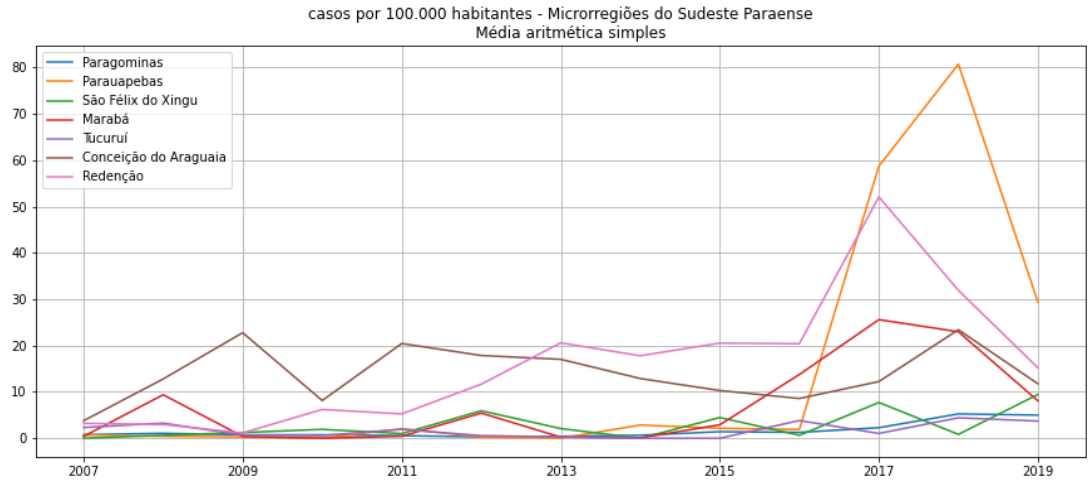
Figura 9 – Evolução da Taxa de Incidência das Mesorregiões Paraenses



Fonte: própria

Ao analisar as microrregiões da mesorregião Sudeste do Pará, observou-se que a microrregião de Parauapebas teve taxas muito altas no final do período analisado, como pode ser observado na Figura 10, sugerindo que ali estão localizados alguns surtos da doença.

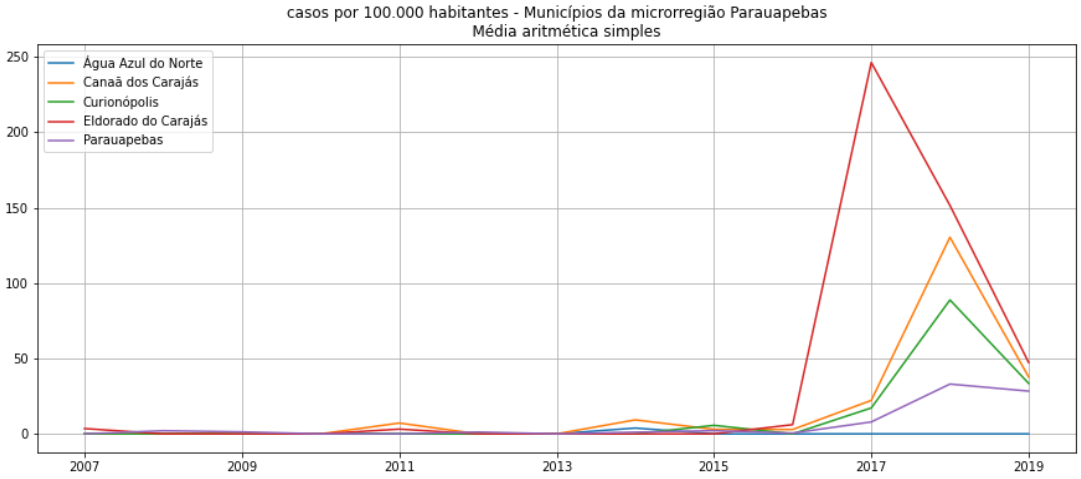
Figura 10 – Evolução da Taxa de Incidência das Microrregiões do Sudeste Paraense



Fonte: própria

Eldorado do Carajás, em 2016, teve aproximadamente 6 casos confirmados para cada 100.000 habitantes, em 2017 este número saltou para 246 conforme apresenta a figura 11.

Figura 11 – Evolução da Taxa de Incidência dos Municípios da Microrregião Parauapebas



Fonte: própria

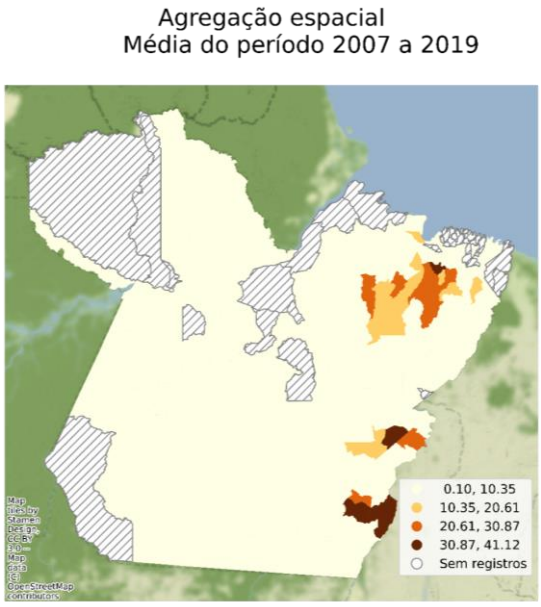
Até este ponto foi feita uma exploração inicial para identificar se havia indícios de agrupamento nos dados em relação à taxa de incidência. Foi constatado que sim e fica ainda mais evidente ao se utilizar algumas visualizações mais avançadas.

Nas figuras 12 e 13, são apresentados dois mapas coropléticos dos municípios paraenses, mostrando agregação espacial, o primeiro com a média de todo o período e o segundo com a média do último triênio. Na figura 12 observa-se que os municípios com as

maiores taxas (cores mais escuras) estão agrupados no nordeste e sudeste do estado. Na figura 13, os locais com piores taxas (média trienal), estão no intervalo de 111 a 148 casos para cada 100.000 habitantes e situados no sudeste do Pará.

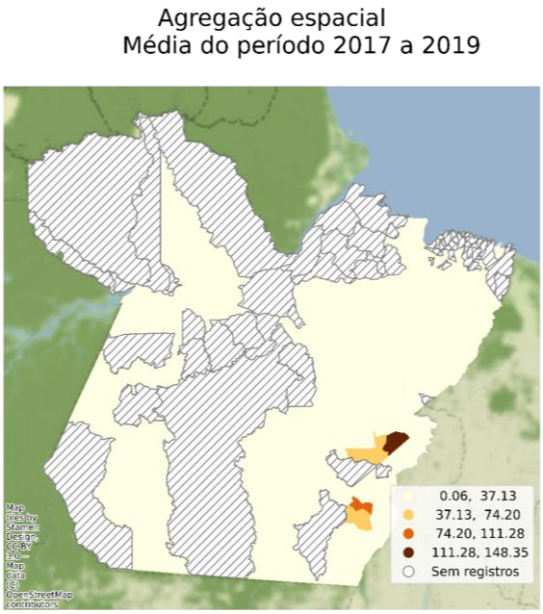
Os mapas indicam de forma mais clara o que foi visualizado anteriormente: no período de estudo os casos se concentram em duas regiões, nordeste e sudeste do referido estado, mas no último triênio – figuras 12 e 13 –, os municípios com piores taxas estão no sudeste do Pará apenas.

Figura 12 – Distribuição Espacial da Taxa de Incidência Média do Período 2007-2019 no Estado do Pará



Fonte: própria

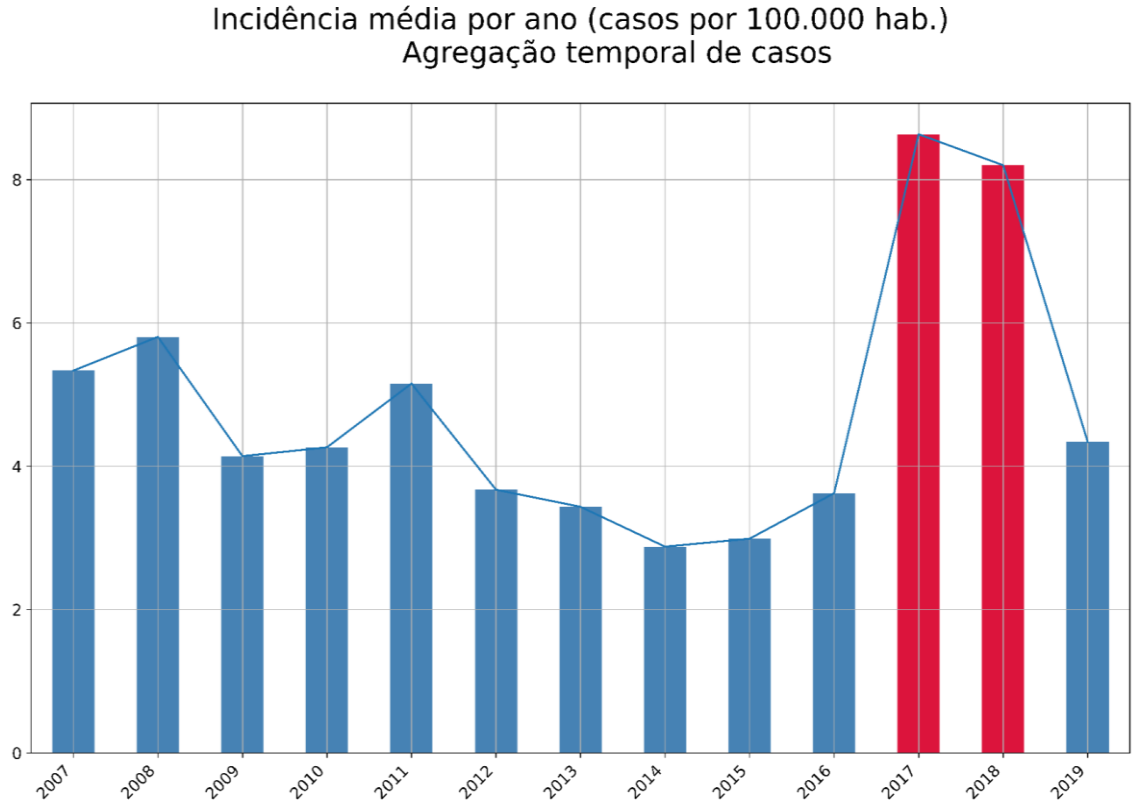
Figura 13 – Distribuição Espacial da Taxa de Incidência Média do Período 2017-2019 no Estado do Pará



Fonte: própria

Ao visualizar as taxas médias anuais de todo o estado, no gráfico de barras da figura 14, observa-se que nos anos 2017 e 2018 houve um forte aumento nas taxas no estado, indicando que houve agregação temporal neste período.

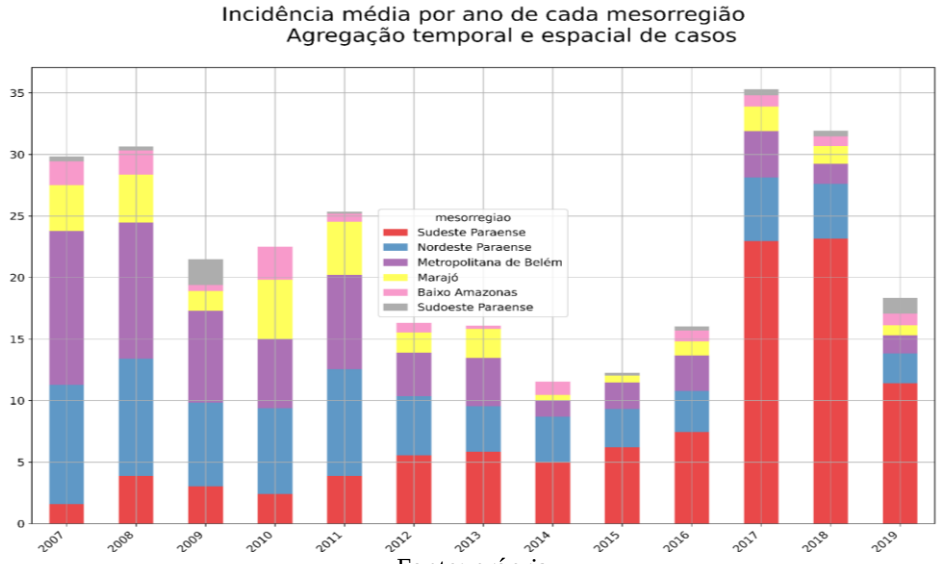
Figura 14 – Distribuição Temporal da Taxa de Incidência Média do Período 2007-2019 no Estado do Pará



Fonte: própria

A agregação espaço-temporal pode ser visualizada pelo gráfico de barras empilhadas na figura 15. Observa-se que as mesorregiões Nordeste do Pará, Metropolitana e Marajó juntas lideram com a maior quantidade de casos por 100.000 habitantes até o ano de 2013. A partir de 2014 os seus casos foram reduzidos e o Sudeste paraense começou a liderar com as piores taxas, vivendo um momento crítico a partir de 2017.

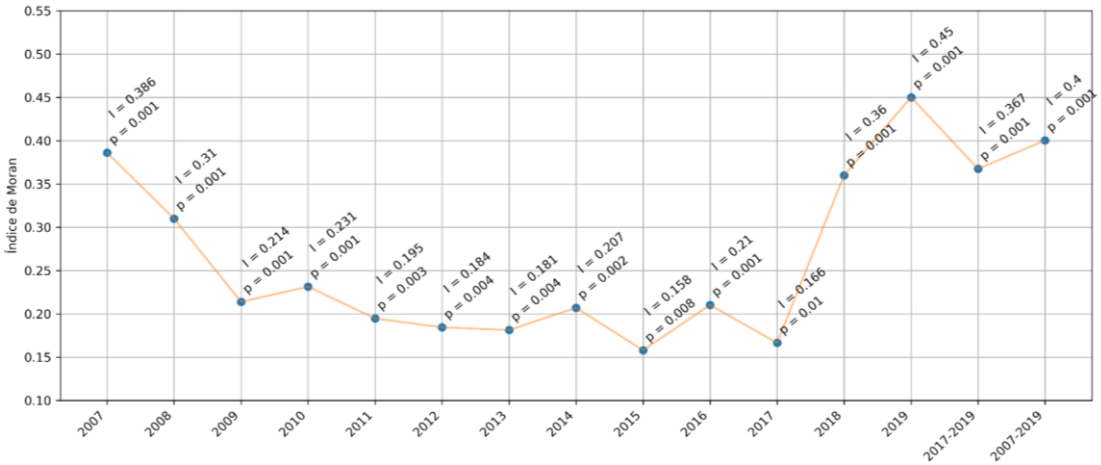
Figura 15 – Distribuição Espaço-Temporal da Taxa de Incidência Média do Período 2007-2019 no Estado do Pará



Fonte: própria

As visualizações foram muito úteis, mas para complementar a análise o Índice de Moran Global foi utilizado para medir a autocorrelação espacial entre os municípios nos treze anos. O Índice foi positivo em todo o período, sendo maior em 2007 e 2019, figura 16, e todos os *p-value* dos testes de pseudo significância, com 999 permutações, foram menores que 0,013.

Figura 16 – Índice de Moran Aplicado à Taxa de Incidência para o Estado do Pará no Período 2007-2019



Fonte: própria

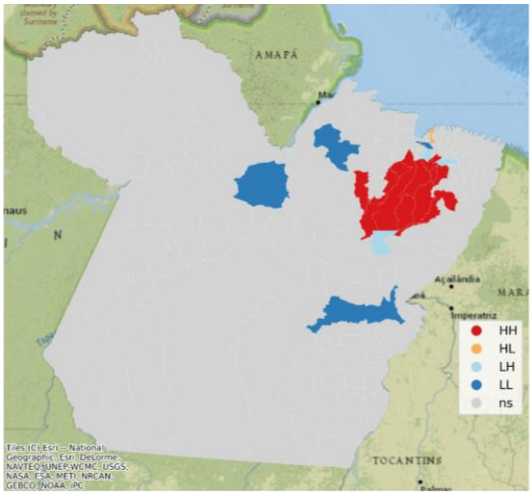
O índice global é uma estatística que representa todo o território, entretanto pode haver alguns locais com taxas mais altas, pontos quentes, e outros com taxas baixas, pontos frios. A

decomposição do Índice de Moran Global pode ser feita através da estatística LISA. O mapa LISA permite identificar padrões de agrupamento, assim, da figura 17 à figura 29 é possível ver o referido mapa para os anos de 2007 a 2019, já a figura 30 apresenta o mapa LISA para o triênio 2017-2019.

Os municípios em vermelho são locais com taxas altas circundados de locais com taxas altas. Os municípios em azul escuro são locais com taxas baixas e próximos de outros locais com taxas baixas. Em azul claro, têm-se municípios com taxas baixas que estão próximos de outros com taxas altas. A região cinza é estatisticamente insignificante e são áreas de transição.

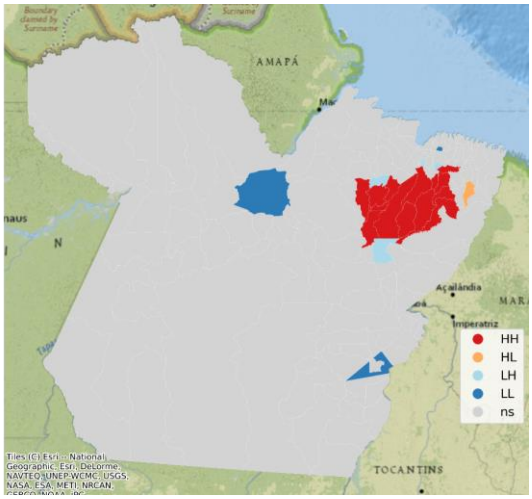
Através dos mapas LISA é possível visualizar a dinâmica da doença no território no período de estudo.

Figura 17 – Mapa LISA no Ano 2007



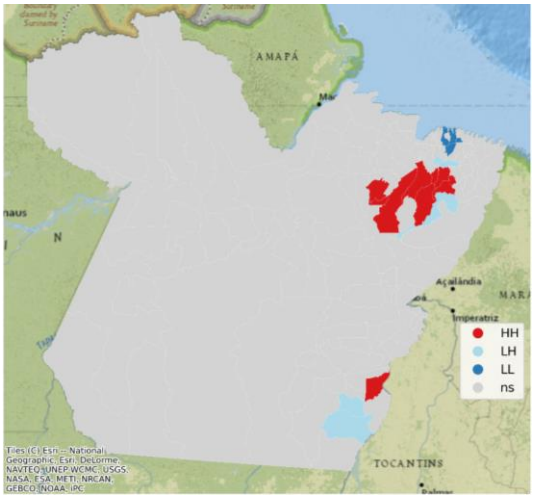
Fonte: própria

Figura 18 – Mapa LISA no Ano 2008



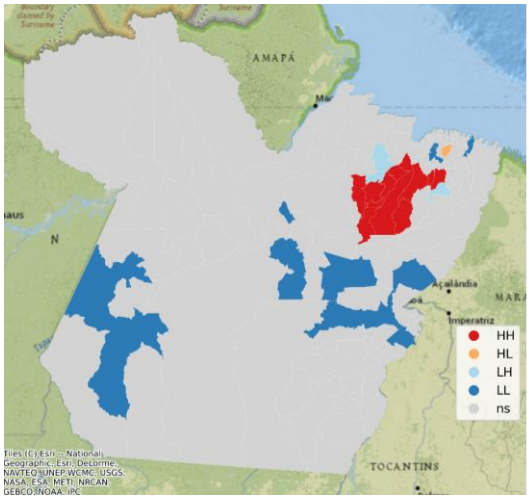
Fonte: própria

Figura 19 – Mapa LISA no Ano 2009



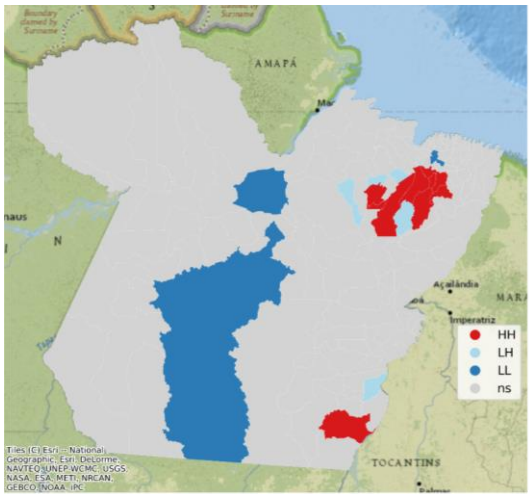
Fonte: própria

Figura 20 – Mapa LISA no Ano 2010



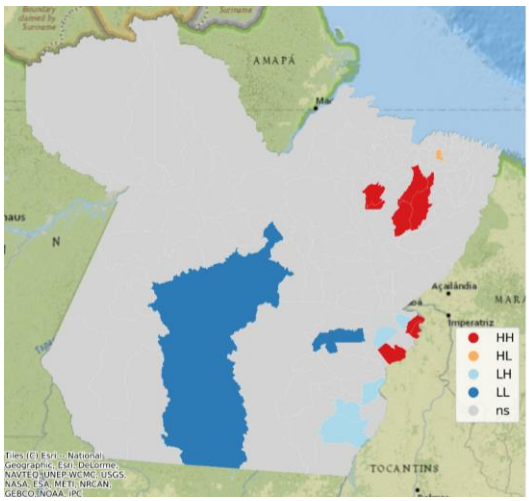
Fonte: própria

Figura 21 – Mapa LISA no Ano 2011



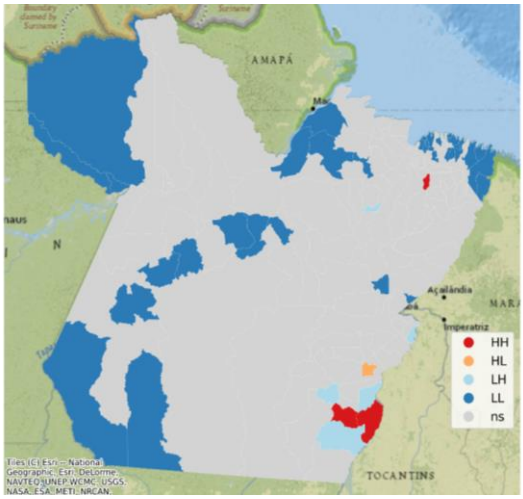
Fonte: própria

Figura 22 – Mapa LISA no Ano 2012



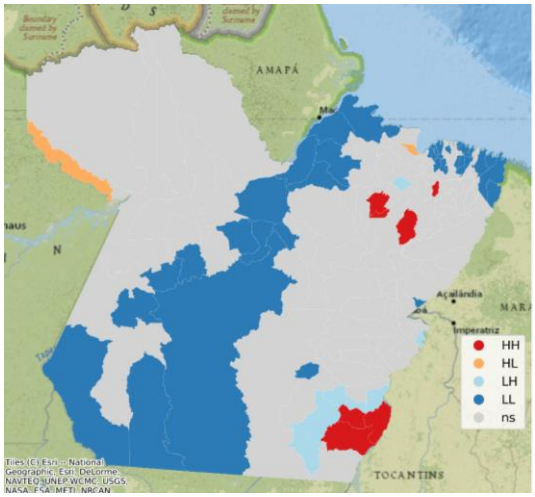
Fonte: própria

Figura 23 – Mapa LISA no Ano 2013



Fonte: própria

Figura 24 – Mapa LISA no Ano 2014



Fonte: própria

Figura 25 – Mapa LISA no Ano 2015

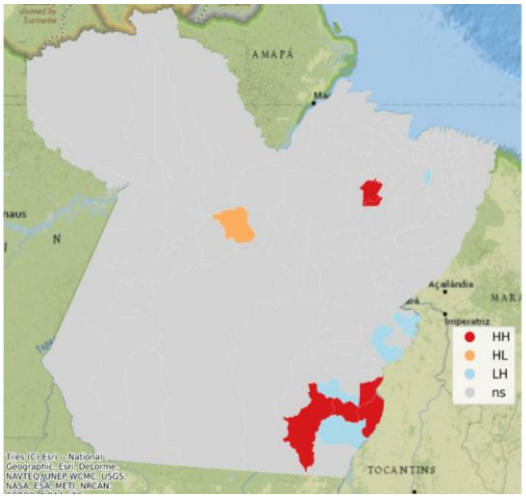
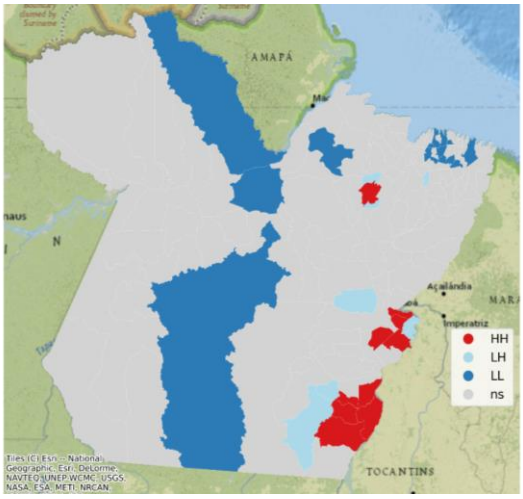


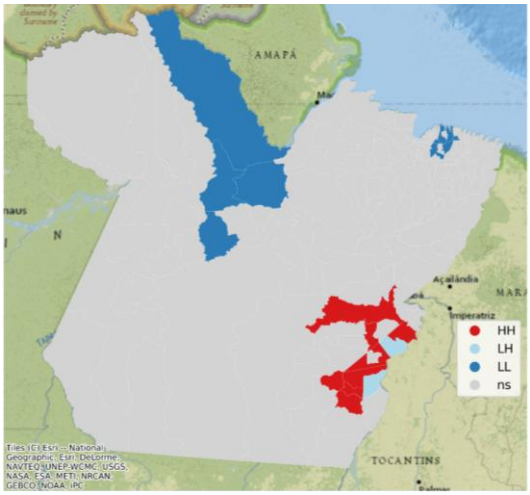
Figura 26 – Mapa LISA no Ano 2016



Fonte: própria

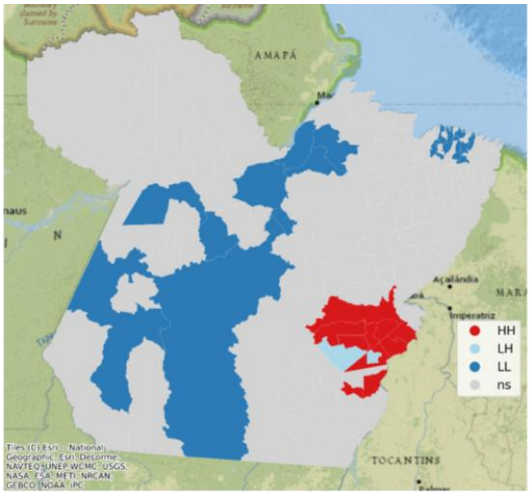
Fonte: própria

Figura 27 – Mapa LISA no Ano 2017



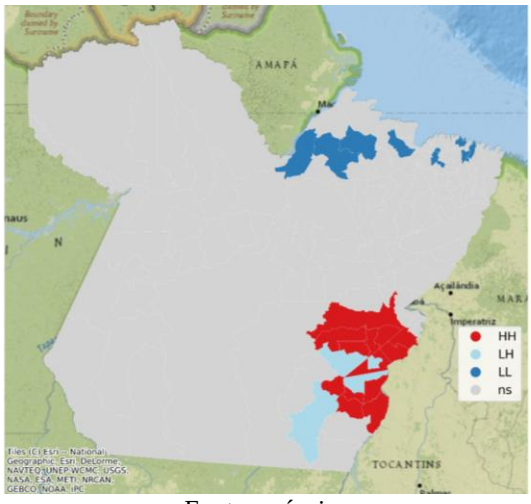
Fonte: própria

Figura 28 – Mapa LISA no Ano 2018



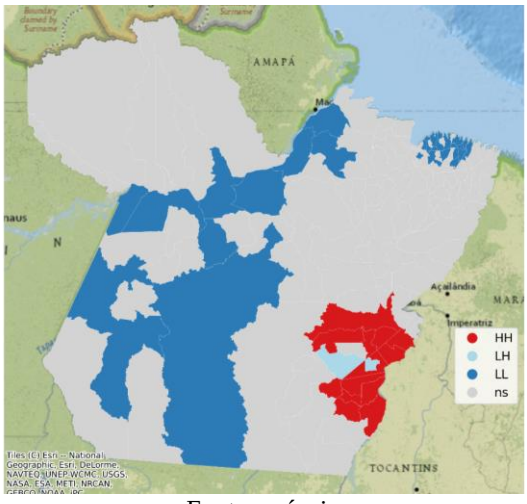
Fonte: própria

Figura 29 – Mapa LISA no Ano 2019



Fonte: própria

Figura 30 – Mapa LISA no Triênio 2017-2019



Fonte: própria

Finalmente, a estatística de varredura espaço-temporal foi aplicada através do *software* SaTScan. A técnica varreu todo o espaço e tempo em busca de aglomerados espaço-temporais e retornou três *clusters* para o período.

O primeiro *cluster* ocorreu entre 2007 e 2013 (7 anos), está localizado na mesorregião do Marajó. O agrupamento tem um raio de aproximadamente 15 quilômetros e incluiu apenas um município, Salvaterra.

Tabela 1 – Resultados do *cluster 1*

Região do Marajó - 2007 a 2013				
População	**Municípios (1 município)**			**Coordenadas**
21.057 hab	Salvaterra			0,725925 S, 48,516013 W) / 14,90 km
Observados	**Esperados**	**Casos anuais / 100.000**	**Observado/Esperado**	**Risco Relativo**
36	5,50	26,4	6,54	6,59

Fonte: própria.

A tabela 2 refere-se ao agrupamento ocorrido entre 2007 e 2013 (7 anos), nas mesorregiões Nordeste e Região Metropolitana de Belém. O agrupamento tem um raio de aproximadamente 176 quilômetros e incluiu 22 municípios.

Tabela 2 – Resultados do *cluster* 2

Nordeste Paraense e Região Metropolitana de Belém - 2007 a 2013				
População	**Municípios (22 municípios)**			**Coordenadas**
1.326.257 hab	Tailândia, Moju, Mocajuba, Tomé-Açu, Baião, Cametá, Igarapé-Miri, Breu Branco, Ipixuna do Pará, Acará, Goianésia do Pará, Oeiras do Pará, Limoeiro do Ajuru, Abaetetuba, Tucuruí, Bagre, Paragominas, Aurora do Pará, Concórdia do Pará, Barcarena, Bujaru e São Domingos do Capim.			(2,937080 S, 48,951282 W) / 175,75 km
Observados	**Esperados**	**Casos anuais / 100.000**	**Observado/Esperado**	**Risco Relativo**
1.356	349,36	15,7	3,88	5,27

Fonte: própria.

O agrupamento mais recente foi identificado na mesorregião Sudeste, entre 2014 e 2019 (6 anos), com raio de aproximadamente 179 quilômetros, incluiu 17 municípios e apresentou o maior risco relativo dentre os agrupamentos encontrados. A tabela 3 resume as informações.

Tabela 3 – Resultados do *cluster 3*

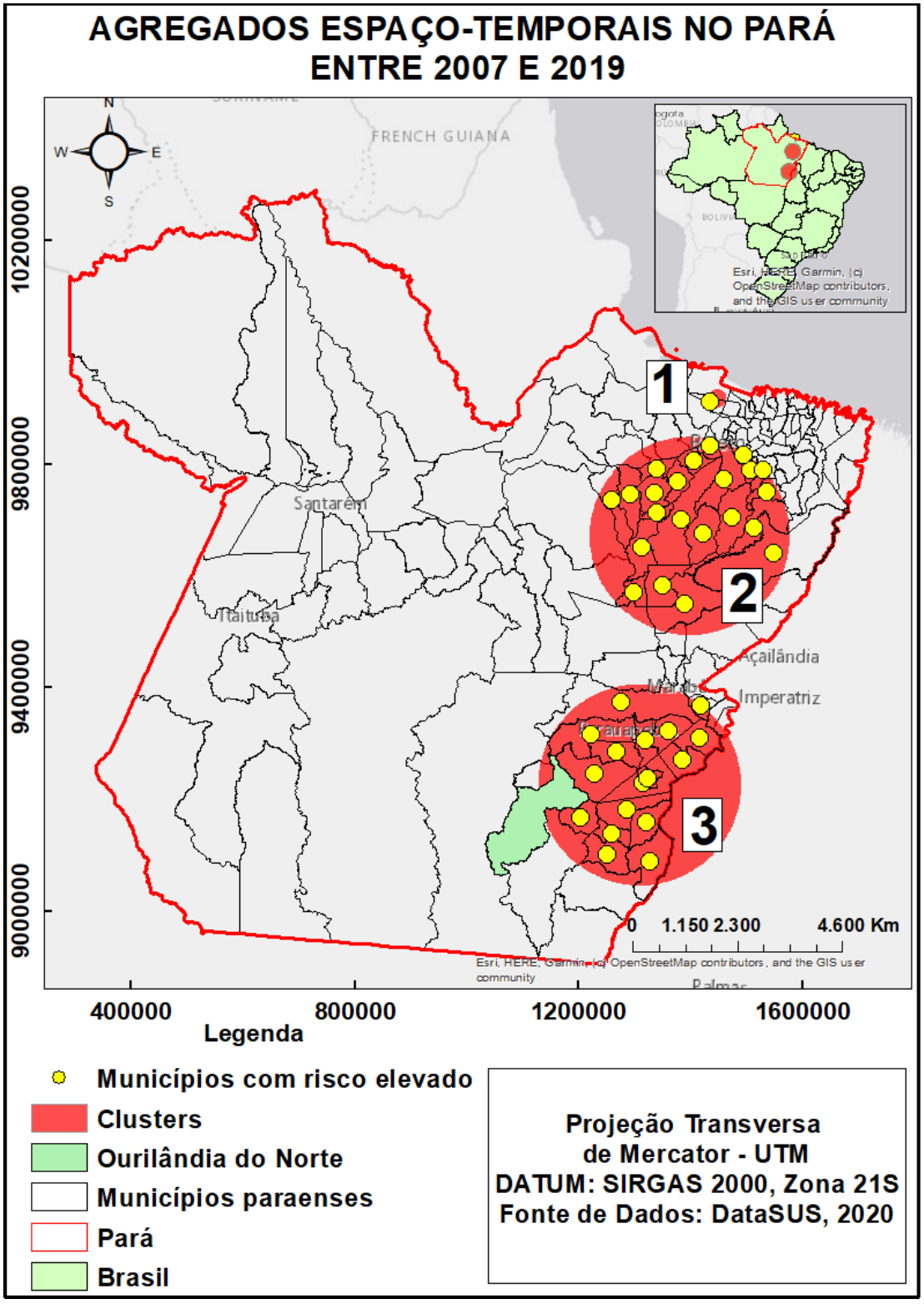
Sudeste Paraense - 2014 a 2019

População	Municípios (17 municípios)			Coordenadas
804.688 hab	Xinguara, Sapucaia, Rio Maria, Floresta do Araguaia, Canaã dos Carajás, Curionópolis, Água Azul do Norte, Piçarra, Pau D'Arco, Eldorado do Carajás, Bannach, Parauapebas, São Geraldo do Araguaia, Redenção, Conceição do Araguaia, Marabá e São Domingos do Araguaia.			(6,942857 S, 49,703581 W) / 178,74 km
Observados	**Esperados**	**Casos anuais / 100.000**	**Observado/Esperado**	**Risco Relativo**
1.256	209,30	24,2	6,00	8,16

Fonte: própria.

A figura 31 apresenta a localização geográfica dos agrupamentos encontrados (círculos em vermelho). Os pontos amarelos representam as sedes dos municípios incluídos nos agrupamentos. A área em verde claro representa o município de Ourilândia do Norte, que não foi adicionado ao *cluster* pelo algoritmo, mas observa-se a proximidade do mesmo e que municípios vizinhos foram incluídos.

Figura 31 – *Clusters* Espaço-Temporais Identificados no Estado do Pará entre os anos 2007 e 2019



Fonte: própria

6 CONCLUSÕES E TRABALHOS FUTUROS

Motivado pelo aumento de casos de LVH no final da década de 2020 no município de Ourilândia do Norte (Pará) e pela constatação de que municípios próximos também enfrentavam situação semelhante, este estudo foi idealizado e executado com o objetivo de avaliar a distribuição da doença no território paraense e verificar se ela estava uniformemente distribuída ou apresentava características de agrupamento, no período 2007 a 2019.

A investigação dos dados de LVH do SINAN concluiu que existiram grupos de municípios com autocorrelação alta (Índice de Moran local), em relação à taxa de incidência, principalmente na parte leste do estado, compreendendo as mesorregiões Nordeste paraense, Marajó, Metropolitana e Sudeste paraense. Este resultado foi confirmado e complementado pela análise de agrupamentos espaço-temporais, identificando três *clusters*, sendo dois no período de 2007 a 2013 (Nordeste Paraense, Região Metropolitana e Marajó) e outro no período 2014 a 2019 (Sudeste do Pará), concluindo que, de fato, houve agrupamento da doença. O agrupamento mais recente (Sudeste), 2014 a 2019, compreende 17 municípios, dentre os quais não está o de Ourilândia do Norte, entretanto foram incluídos municípios vizinhos.

Este estudo não teve como objetivo explicar ou indicar causas da expansão da doença no território, visto que exige conhecimento especialista da área, por exemplo da epidemiologia. A investigação de causas, variáveis associadas e posterior intervenção são muito importantes para a contenção eficiente da doença em uma região.

A revisão bibliográfica apontou que são vários os fatores que podem influenciar no aumento de casos de LVH. Por ser uma doença vetorial, fatores socioambientais e climáticos influenciam no aumento de vetores ou aproximação destes com os seres humanos. Mas, para que haja a transmissão é necessário que o vetor esteja infectado, e isto está relacionado com a presença de reservatórios da doença no ambiente. Também, apesar de o cão doméstico ser o principal reservatório da doença no Brasil, animais silvestres também o são. Os reservatórios e vetores também podem se movimentar e levar a doença para outros locais. Logo, o aparecimento e manutenção da doença em um local podem estar associados a vários fatores e a identificação destes pode contribuir para uma intervenção mais eficiente.

A coleta e investigação de variáveis associadas à LVH nos agrupamentos identificados neste trabalho pode produzir resultados interessantes, pode-se inferir que os fatores mais associados sejam os mesmos ou diferentes nos diferentes *clusters*. Também, a inclusão de áreas fronteiriças ao estado do Pará nesta análise pode ajudar a esclarecer se a doença está migrando

de estados vizinhos. Ressalta-se que a presença de agrupamentos avançando no tempo de outros estados para o Pará não são suficientes para provar esta relação.

Por fim, estas são algumas possibilidades de extensão desta pesquisa e ficam como sugestões para trabalhos futuros.

REFERÊNCIAS

AHMAD, A. Epidemiology and spatiotemporal analysis of visceral leishmaniasis in Palestine from 1990 to 2017. **International Journal of Infectious Diseases**, v. 90, p. 206–212, 1 jan. 2020.

ANSELIN, L. Local Indicators of Spatial Association – LISA. **Geographical Analysis**, v. 27, n. 2, p. 93–115, abr. 1995.

ARAÚJO, D. DA C. Análise espacial dos casos humanos de leishmaniose visceral. **Arquivos de Ciências da Saúde**, v. 24, n. 2, p. 71–75, 5 jul. 2017.

BIRANT, D.; KUT, A. ST-DBSCAN: An algorithm for Clustering Spatial–Temporal Data. **Data & Knowledge Engineering**, Intelligent Data Mining. [*s. l.*], v. 60, n. 1, p. 208-221, jan. 2007.

BRASIL. Ministério da Saúde. Gabinete do Ministro. **Portaria nº 204, de 17 de fevereiro de 2016.** Define a Lista Nacional de Notificação Compulsória de doenças, agravos e eventos de saúde pública nos serviços de saúde públicos e privados em todo o território nacional, nos termos do anexo, e dá outras providências. [*S. l.*], 17 fev. 2016. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2016/prt0204_17_02_2016.html?fbclid=IwAR15gO94x14EALuvcAK3-wfBNIBO65lxA0WvpZ_1WYh-lImVa3PsfTnZ5w4. Acesso em: 7 mar. 2022.

BRASIL. Ministério da Saúde. **Leishmaniose Visceral:** Instruções para preenchimento da Ficha de Investigação, 27 set. 2006. Disponível em: http://portalsinan.saude.gov.br/images/documentos/Agravos/Leishmaniose%20Visceral/LV_v5_instr.pdf. Acesso em: 8 mar. 2022.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Coordenação-Geral de Desenvolvimento da Epidemiologia em Serviços. **Guia de Vigilância em Saúde.** 1. ed. atual. Brasília: Ministério da Saúde, 2016.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Coordenação-Geral de Desenvolvimento da Epidemiologia em Serviços. **Guia de Vigilância em Saúde:** volume único. 3. ed. Brasília: Ministério da Saúde, 2019. Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/guia_vigilancia_saude_3ed.pdf. Acesso em: 7 mar. 2022.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância de Doenças Transmissíveis. Coordenação-Geral de Doenças Transmissíveis. **Caderno de Indicadores:** Leishmaniose Tegumentar Leishmaniose Visceral, 2018. Disponível em: http://portalsinan.saude.gov.br/images/documentos/Agravos/LTA/Indicadores_Leishmanioses_2018.pdf. Acesso em: 8 mar. 2022.

CARDIM, M. F. M. *et al*. Leishmaniose Visceral no Estado de São Paulo, Brasil: Análise Espacial e Espaço-Temporal. **Revista de Saúde Pública**, v. 50, 2016.

CARVALHO, A. C. *et al*. Leishmaniose Visceral Humana: Análise da Série Histórica em

Âmbito Nacional e Local. *In*: **Anais do IV Congresso de Educação e Saúde do Sudeste do Pará**. Marabá: Universidade do Estado do Pará, 2019. Disponível em: https//www.even3.com.br/anais/CONESP/221060-LEISHMANIOSE-VISCERAL-HUMANA--ANALISE-DA-SERIE-HISTORICA-EM-AMBITO-NACIONAL-E-LOCAL>. Acesso em: 14 mar. 2022.

CHENG, T. *et al.* Spatiotemporal Data Mining. *In*: FISCHER, M. M.; NIJKAMP, P. (eds.)., **Handbook of Regional Science**. Berlin, Heidelberg: Springer, 2014. p. 1173-1193.

DRUCK, S. *et al.* Análise Espacial de Áreas. *In*: DRUCK, S. *et al.* **Análise Espacial de Dados Geográficos**. Brasília, EMBRAPA, 2004.

FONTOURA, I. G.; FONTOURA, V. M.; NASCIMENTO, L. F. C. Spatial analysis of the occurrence of visceral leishmaniasis in the state of Tocantins, Brazil. **Ambiente e Agua - An Interdisciplinary Journal of Applied Science**, v. 11, n. 5, p. 1088-1095, 10 dez. 2016.

FURTADO, A. S. *et al.* Space-time analysis of visceral leishmaniasis in the State of Maranhão, Brazil. **Ciência & Saúde Coletiva**, v. 20, n. 12, p. 3935–3942, dez. 2015.

GALGAMUWA, L. S.; DHARMARATNE, S. D.; IDDAWELA, D. Leishmaniasis in Sri Lanka: spatial distribution and seasonal variations from 2009 to 2016. **Parasites & Vectors**, v. 11, n. 1, p. 60, 25 jan. 2018.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining**: um Guia Prático. 1. ed. Rio de Janeiro: Elsevier, 2005.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques**. 3. ed. Burlington, MA: Elsevier, 2012.

HAO, Y. *et al.* Spatio-temporal clustering of Mountain-type Zoonotic Visceral Leishmaniasis in China between 2015 and 2019. **PLOS Neglected Tropical Diseases**, v. 15, n. 3, p. e0009152, 2021.

IBGE. Instituto Brasileiro de Geografia e Estatística. IBGE - Cidades: Pará, História e Fotos, 2017. Disponível em: https://cidades.ibge.gov.br/brasil/pa/historico. Acesso em: 14 jan. 2022.

JACQUEZ, G. M. A K Nearest Neighbor Test for Space-Time Interaction. **Statistics in Medicine**, v. 15, n. 18, p. 1935–1949, 1996.

KULLDORFF, M. **SaTScan Manual do Usuário**: Versão do Manual Traduzido para o Português. mai. 2016. Disponível em: https://www.satscan.org/SaTScan_TM_Manual_do_Usu%C3%A1rio_v9.4_Portugues.pdf. Acesso em: 20 ago. 2021.

MACHADO, G. *et al.* Revisiting area risk classification of visceral leishmaniasis in Brazil. **BMC Infectious Diseases**, v. 19, n. 1, p. 2, 3 jan. 2019.

MAIMON, O.; ROKACH, L. Introduction to Knowledge Discovery and Data Mining. *In*: MAIMON, O.; ROKACH, L. **Data Mining and Knowledge Discovery Handbook**. 2. ed. New York, USA: Springer, 2010. p. 1-15.

OLIVEIRA, M. V. M. de *et al.* Estudo Epidemiológico dos Casos Confirmados de Leishmaniose Visceral no Município de Marabá no Período de 2013 a 2017. In: **IV Congresso de Educação e Saúde do Sudeste do Pará**. Marabá, Pará: Universidade do Estado do Pará, 2019. Disponível em: https://www.even3.com.br/anais/CONESP/221117. Acesso em: 22 abr. 2021.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. **Leishmanioses: Informe epidemiológico nas Américas**. Núm. 9. Washington, D.C.: OPAS, 3 dez. 2020. Disponível em: https://iris.paho.org/handle/10665.2/53091. Acesso em: 22 abr. 2021.

PETRUZALEK, D. READ. DBC: Um Pacote para Importação de Dados do DataSUS na Linguagem R. *In*: **Anais do XV Congresso Brasileiro de Informática em Saúde**. S. l.: s. n., 2016. p. 27-30.

SILVA, W. J. *et al.* Spatiotemporal patterns and integrated approach to prioritize areas for surveillance and control of visceral leishmaniasis in a large metropolitan area in Brazil. **Acta Tropica**, v. 211, p. 105615, nov. 2020.

SILVEIRA, F. T. *et al.* Revendo a trajetória da leishmaniose visceral americana na Amazônia, Brasil: de Evandro Chagas aos dias atuais. **Revista Pan-Amazônica de Saúde**, v. 7, n. esp., p. 15–22, dez. 2016.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. 1. ed. Harlow, UK: Pearson, 2014.

WALLER, L. A.; GOTWAY, C. A. **Applied Spatial Statistics for Public Health Data**. Hoboken, N.J: John Wiley & Sons, 2004.

WERNECK, G. L.; STRUCHINER, C. J. Estudos de Agregados de Doença no Espaço-Tempo: Conceitos, Técnicas e Desafios. **Cadernos de Saúde Pública**, v. 13, p. 611–624, out. 1997.

ZHENG, C. *et al.* Visceral leishmaniasis in northwest China from 2004 to 2018: a spatio-temporal analysis. **Infectious Diseases of Poverty**, v. 9, n. 1, p. 165, 3 dez. 2020.