

# Clusterização Hierárquica de Mesorregiões com Base de Dados do SINAN: Um Estudo sobre Tuberculose no Estado de Pernambuco

*Hierarchical Clustering of Mesoregions Using SINAN Database: A Study on Tuberculosis in Pernambuco State*

**Franciele França<sup>1</sup>**

[orcid.org/0009-0004-1504-3473](https://orcid.org/0009-0004-1504-3473)

**Roberta Fagundes<sup>2</sup>**

[orcid.org/0000-0002-7172-4183](https://orcid.org/0000-0002-7172-4183)

<sup>1</sup>Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: [fsf2@poli.br](mailto:fsf2@poli.br)

<sup>2</sup>Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: [roberta.fagundes@upe.br](mailto:roberta.fagundes@upe.br)

**DOI: 10.25286/repa.v11i3.3928**

Esta obra apresenta Licença Creative Commons Atribuição-Não Comercial 4.0 Internacional.

## RESUMO

A tuberculose permanece como desafio relevante de saúde pública no Brasil, com Pernambuco entre os estados de maior incidência. Este estudo analisou padrões epidemiológicos da doença entre 2001 e 2024, aplicando clusterização hierárquica às mesorregiões com dados do SINAN. O método de Ward, validado por métricas estatísticas e complementado pela PCA, identificou quatro agrupamentos robustos. A Região Metropolitana do Recife destacou-se como epicentro persistente, enquanto demais mesorregiões migraram de baixa/média para incidência intermediária. Observou-se predominância masculina (61–70%) em todos os clusters, especialmente nos de maior carga, além de diferenças etárias relevantes. A clusterização hierárquica mostrou-se vantajosa por revelar agrupamentos em múltiplos níveis, oferecendo interpretações mais ricas e subsidiando estratégias de vigilância adaptadas ao perfil epidemiológico regional.

**PALAVRAS-CHAVE:** Inteligência Artificial; Clusterização Hierárquica; Tuberculose; Saúde Pública;

## ABSTRACT

Tuberculosis remains a major public health concern in Brazil, with Pernambuco among the states of highest incidence. This study applied hierarchical clustering to epidemiological data from SINAN (2001–2024) to identify regional patterns. Ward's method, validated through Silhouette, Calinski-Harabasz, and Davies-Bouldin indices, revealed four robust clusters. The Metropolitan Region of Recife consistently showed high incidence, while other mesoregions transitioned from low to intermediate levels. Male predominance (61–70%) was observed across clusters, especially in high-incidence groups. Hierarchical clustering proved advantageous by uncovering multi-level groupings, offering richer interpretability than partitioning methods, and supporting tailored surveillance strategies in public health planning.

**KEYWORDS:** Artificial Intelligence; Hierarchical Clustering; Tuberculosis; Public Health;

## 1 INTRODUÇÃO

A Revista de Engenharia e Pesquisa Aplicada (REPA) é um meio de divulgação científica para trabalhos inéditos das áreas de Engenharias, Computação e tecnológicas afins como: inteligência artificial, ciência de dados, mineração de dados e saúde pública. É utilizado pela REPA o sistema de editoração aberta *Open Journal System* (OJS).

A tuberculose, causada pela bactéria *Mycobacterium tuberculosis* (bacilo de Koch), é uma doença infecciosa e transmissível. Ela afeta principalmente os pulmões, podendo também atingir outros órgãos, caracterizando a forma extrapulmonar [1]. No cenário mundial, ela ressurgiu como uma das principais causa de morte, superando a COVID-19. A Organização Mundial da Saúde (OMS) incluiu o Brasil entre os países prioritários para o combate à doença [2].

Dados do Ministério da Saúde revelam que, após uma queda histórica durante a pandemia, os casos voltaram a crescer. Em 2024, foram registradas mais de 85 mil notificações e mais de 6 mil óbitos anuais. O estado de Pernambuco figura entre as dez unidades federativas com maior coeficiente de incidência. A taxa foi de 55,2 casos por 100 mil habitantes em 2024 [2].

Diante desse panorama, é de extrema importância a realização de análises de padrões epidemiológicos, para compreender a dinâmica da tuberculose e subsidiar estratégias de vigilância e controle. Este artigo tem como objetivo aplicar técnicas de clusterização hierárquica às mesorregiões do estado de Pernambuco, utilizando dados do Sistema de Informação de Agravos de Notificação (SINAN), de modo a identificar agrupamentos regionais e padrões epidemiológicos da doença. Devido à organização territorial, de Pernambuco, composta por apenas cinco mesorregiões, o método de clusterização é especialmente adequado para o estudo. Esta técnica não necessita de um grande volume de unidades espaciais para ser eficiente. Desta forma, os clusters podem revelar mesorregiões com maiores incidências ou taxas mais homogêneas, áreas com surtos concentrados, regiões que compartilham determinantes socioeconômicos ou ambientais. Revelando padrões que podem auxiliar gestores e pesquisadores na vigilância e no planejamento de ações de saúde.

A seção 2, apresenta a revisão de literatura, abordando estudos sobre tuberculose e aplicações de inteligência artificial, como clusterização, em diversas problemáticas da sociedade. Dando prosseguimento, a seção 3, descreve a metodologia utilizada, incluindo a base de dados, os critérios de análise e o método de clusterização hierárquica. A discussão dos resultados obtidos e suas implicações para a vigilância epidemiológica estão na seção 4. Por fim, na seção 5, apresentam-se as conclusões, destacando as contribuições do estudo e possíveis direções para pesquisas futuras.

## 2 REVISÃO DA LITERATURA

Estudos realizados na Região Metropolitana do Recife (RMR), entre 2019 e 2020 evidenciaram a elevada concentração de casos, sobretudo em homens jovens, com impacto significativo na mortalidade. Foram registrados 11.678 casos no estado de Pernambuco, dos quais 69,78% ocorreram na região metropolitana, com destaque na capital Recife com 18,26% do total. O trabalho confirma a vulnerabilidade da população ativa e a necessidade de medidas de vigilância e controle direcionada [3].

No campo da análise de dados, diferentes abordagens de mineração de dados e aprendizado de máquina têm sido aplicadas para compreender padrões epidemiológicos da doença. No Brasil, um estudo utilizando K-Means em séries temporais do SINAN (2001-2022) identificou três agrupamentos de estados com diferentes níveis de prevalência, demonstrando a utilidade da clusterização para detectar tendências temporais e regionais [4]. Outras pesquisas internacionais reforçam essa perspectiva. Um estudo realizado em Makassar, a integração de K-Means com Sistemas de Informação Geográfica (SIG) permitiu visualizar a dispersão espacial da tuberculose considerando fatores socioeconômicos [5].

Na província de Java Oriental, o uso do algoritmo K-Medoids possibilitou mapear diferentes vulnerabilidades, destacando a importância de técnicas alternativas de clusterização [6].

Além disso, diversos estudos publicados anteriormente na REPA também demonstram a relevância da aplicação de técnicas de mineração de dados e clusterização em diferentes contextos. No campo da saúde pública, destaca-se o trabalho de

Alves et al. [7], que realizou um estudo comparativo de técnicas de previsão para casos de dengue na cidade do Recife, utilizando a abordagem CRISP-DM. O estudo reforçou a importância de modelos preditivos para apoiar políticas de vigilância epidemiológica.

Em outro contexto, Rocha et al. [8] aplicaram algoritmos de clusterização para a segmentação de clientes em uma fintech brasileira, comparando K-Means, Fuzzy C-Means e K-Medoids. Foi demonstrado que o K-Medoids, com distância de Gower, apresentou melhor desempenho na identificação de perfis de consumidores, evidenciando o potencial da clusterização para análise de grandes bases de dados.

Ainda nos artigos publicados anteriormente na REPA, Borges et al. [9] exploraram diferentes técnicas de clusterização incluindo K-Means, Agglomerative Clustering e Spectral Clustering. Com objetivo de identificar grupos em uma base de dados derivada do questionário Cannabis Use Disorder Identification Test (CUDIT). Os resultados mostraram que tais técnicas foram capazes de separar grupos com padrões distintos, confirmando a aplicabilidade da clusterização em cenários de saúde e comportamento.

Esses trabalhos reforçam que a clusterização é uma ferramenta versátil e poderosa, com aplicabilidade tanto em epidemiologia como em setores sociais e econômicos. No entanto, observa-se que a maioria das pesquisas se concentra em algoritmos particionais, como K-Means, que exigem a definição prévia do número de grupos. O presente estudo diferencia-se ao propor a aplicação da clusterização hierárquica em dados epidemiológicos da tuberculose em Pernambuco, permitindo identificar agrupamentos regionais em diferentes níveis de agregação e apoio em decisões estratégias de vigilância.

### 3 METODOLOGIA

A metodologia adotada no estudo seguiu uma abordagem de etapas sequenciais, visando garantir consistência na análise e reprodutibilidade dos resultados. Inicialmente, procedeu-se à coleta e preparação dos dados provenientes do Sistema de Informação de Agravos de Notificação (SINAN). Em seguida, foram aplicadas técnicas de tratamento e agregação estatística, permitindo a construção de indicadores epidemiológicos por mesorregião e por ano. Na etapa seguinte, os dados foram submetidos a processos de padronização e agrupamento por meio da clusterização hierárquica, utilizando o

método de Ward. A determinação do número ótimo de clusters foi realizada por meio de diferentes métricas de validação. E posteriormente confirmada por meio da Análise de Componentes Principais (PCA), assegurando a coerência dos padrões identificados.

#### 3.1 FONTE DE DADOS

Foram analisados dados públicos do SINAN, disponibilizados pelo DATASUS [10], referentes ao período de 2001 a 2024. O ano de 2024 foi o último ano disponível na base no momento da coleta. A base de dados foi carregada em formato CSV e filtrado para o estado de Pernambuco. Existem cinco mesorregiões na unidade federativa pernambucana. Elas são, Região Metropolitana do Recife (RMR), Mata Pernambucana, Agreste Pernambucano, Sertão Pernambucano e São Francisco Pernambucano. Registros fora do intervalo alvo (2001-2024) foram removidos. Além disso, foram utilizados arquivos auxiliares contendo a relação de códigos e municípios brasileiros e a relação de unidades federativas, com o objetivo de tornar os atributos mais legíveis e facilitar a interpretação dos resultados.

#### 3.2 TRATAMENTO E PREPARAÇÃO DOS DADOS

Realizou-se tratamento de dados faltantes e inconsistentes. Foi criado um mapeamento de mesorregiões e microrregiões, permitindo identificar a mesorregião de cada município. A cobertura de registros com município válido foi de 100%. Na base de dados final, foram selecionadas as variáveis: mesorregião, município, ano de notificação, ano de nascimento e sexo. As colunas NU\_ANO e ANO\_NASC foram convertidas para inteiros, e a idade dos pacientes foi calculada pela diferença entre ano de notificação e ano de nascimento.

#### 3.3 AGREGAÇÃO ESTATÍSTICA

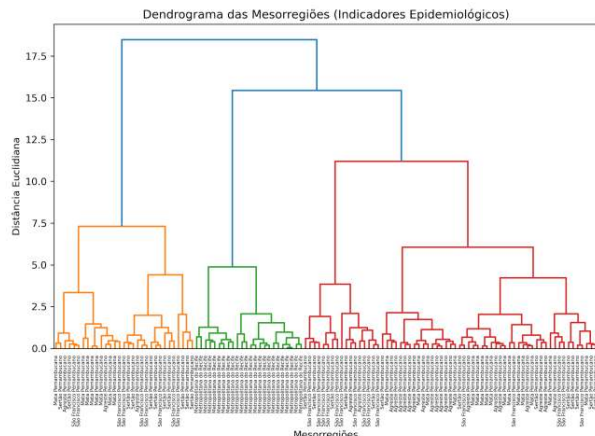
Os dados foram agregados por mesorregião e ano, resultando em indicadores epidemiológicos: idade média dos casos, proporção de homens, proporção de mulheres e número total de casos. Essa agregação permitiu a construção da base final para análise temporal e espacial.

#### 3.4 PADRONIZAÇÃO E CLUSTERIZAÇÃO

## Clusterização Hierárquica de Mesorregiões com Base de Dados do SINAN: Um Estudo sobre Tuberculose no estado de Pernambuco

As variáveis numéricas foram submetidas ao processo de padronização pelo método *z-score*, resultando em uma média igual a zero e desvio padrão igual a um. Esse procedimento evita que diferenças de escala entre variáveis distorçam o processo de agrupamento. Em seguida, aplicou-se o método hierárquico de Ward, que minimiza a variância intra-cluster. O dendrograma (Figura 1) evidencia o agrupamento progressivo das mesorregiões, com base nos indicadores epidemiológicos provenientes do SINAN, revelando grupos distintos de similaridade entre as regiões. A utilização da clusterização hierárquica favorece a identificação de padrões epidemiológicos comuns, destacando mesorregiões que compartilham características semelhantes de incidência, distribuição ou perfil de agravos. Essa vantagem proporcionada por esta técnica permite um direcionamento estratégico de intervenção, monitoramento e prevenção de forma mais precisa e alinhada às particularidades de cada agrupamento identificado.

**Figura 1** – Dendrograma das mesorregiões (indicadores epidemiológicos)



**Fonte:** Os Autores.

### 3.5 DETERMINAÇÃO DO NÚMERO DE CLUSTERS

Para definir o número ótimo de agrupamentos, foram avaliados valores de *k* no intervalo de 2 a 7. A validação interna da clusterização hierárquica foi conduzida por meio de três métricas amplamente utilizadas na literatura: **Silhouette**, **Calinski-Harabasz** e **Davies-Bouldin**. O coeficiente

**Silhouette** mede simultaneamente a coesão intra-cluster e a separação inter-cluster. O índice **Calinski-Harabasz**, avalia a razão entre a dispersão intergrupos e intragrupos. Valores mais elevados refletem clusters mais compactos e bem discriminados. Já o índice **Davies-Bouldin** avalia a similaridade entre clusters, quanto menor o valor, melhor. Os resultados obtidos para cada métrica são descritos na Tabela 1, permitindo comparar o desempenho de diferentes valores de *k* e identificar a configuração mais adequada.

**Tabela 1** - Resultados das métricas de validação para diferentes valores de *k*

| VALOR DO K | SILHOUETTE | CALINSKI-HARABAS Z | DAVIES-BOULDIN |
|------------|------------|--------------------|----------------|
| 2          | 0.358      | 65.2               | 0.993          |
| 3          | 0.440      | 89.2               | 0.822          |
| 4          | 0.455      | 107.0              | 0.730          |
| 5          | 0.441      | 108.3              | 0.770          |
| 6          | 0.401      | 110.1              | 0.839          |
| 7          | 0.394      | 109.6              | 0.823          |

**Fonte:** Os autores.

Os resultados da validação interna indicaram que *k*=4 constituiu a solução mais estatisticamente robusta entre as configurações avaliadas. Esse valor apresentou o maior coeficiente de **Silhouette** (0,455), indicando elevada separabilidade os grupos inter-clusters e coesão intra-cluster. Simultaneamente, o índice de **Davies-Bouldin** (0,730) foi o menor entre todas as alternativas, refletindo baixa razão entre dispersão interna e proximidade intercluster. Característica essencial para a obtenção de grupos compactos e bem discriminados. Além disso, o índice de **Calinski-Harabasz** (107,0), alcançou um valor elevado. Indicando que a variabilidade intergrupos supera de forma expressiva a variabilidade intragrupo. Assim, quatro agrupamentos foram considerados ótimos, fornecendo a melhor separação e coesão entre os perfis epidemiológicos das mesorregiões,

maximizando a estabilidade estatística do modelo e representando de forma mais fiel os padrões epidemiológicos presentes no estado de Pernambuco.

### 3.6 VALIDAÇÃO COM PCA

Para complementar a análise de clusterização hierárquica, empregou-se a Análise de Componentes Principais (PCA) (Figura 2), com o objetivo de projetar o conjunto multidimensional de indicadores epidemiológicos em um espaço bidimensional que preservasse, de forma otimizada, a variabilidade total dos dados. O PCA transforma as variáveis originais em componentes ortogonais que capturam a máxima variância explicada. No presente estudo, os dois primeiros componentes concentraram uma parcela significativa da variabilidade total, assegurando a representatividade da redução dimensional. Essa técnica contribui para a diminuição de colinearidades entre os indicadores, a eliminação de redundâncias e a verificação da coerência dos agrupamentos previamente obtidos quando projetados em um espaço de menor dispersão. A distribuição dos pontos no gráfico de dispersão evidenciou que as mesorregiões atribuídas a um mesmo cluster permaneceram agrupadas no plano dos componentes principais, indicando baixa sobreposição entre grupos e sugerindo separabilidade estatística consistente. Dessa forma, o PCA confirmou a consistência da formação dos grupos e reforçou a robustez da escolha de  $k=4$ .

**Figura 2** – Clusters epidemiológicos nas mesorregiões (PCA)



Fonte: Os Autores.

## 4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

A clusterização hierárquica das mesorregiões de Pernambuco evidenciou padrões epidemiológicos heterogêneos da tuberculose entre 2001 e 2024. Essa heterogeneidade, observada tanto no espaço quanto no tempo, confirma a complexidade da

dinâmica da doença e reforça a necessidade de estratégias de vigilância adaptadas para as características de cada área. Os achados dialogam com estudos anteriores que apontam a influência de fatores demográficos, sociais e econômicos na distribuição da tuberculose, como evidenciado em pesquisas realizadas no Brasil [3;4;11] e em análises internacionais que aplicaram algoritmos de clusterização para compreender a evolução da doença [5;6;12].

### 4.1 RESULTADOS

A aplicação do método de clusterização hierárquica identificou quatro agrupamentos distintos de mesorregiões em Pernambuco no período de 2001 a 2024 descritos na Figura 3 e Quadro 1.

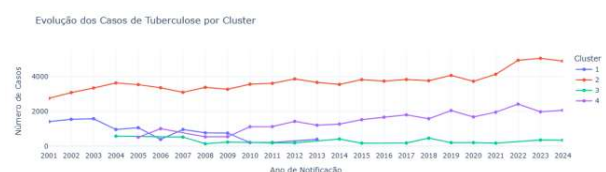
O cluster 1, esteve presente apenas entre 2001 e 2013, indicando áreas que perderam relevância epidemiológica ou foram absorvidas por outros perfis.

No cluster 2, ele apresentou um crescimento contínuo ao longo dos anos, com pico em 2023 (5.059 casos) e leve queda em 2024 (4.898), representando regiões críticas, com maior carga de tuberculose.

Já o cluster 3, iniciou em 2004 e persistiu até 2024, sempre com os menores valores, representando regiões de baixa incidência.

E por último, o cluster 4, surgiu em 2005 e manteve trajetória próxima ao cluster 2, caracterizando regiões de incidência intermediária.

**Figura 3** – Evolução dos casos de tuberculose por cluster (2001-2024)



Fonte: Os Autores.

## Clusterização Hierárquica de Mesorregiões com Base de Dados do SINAN: Um Estudo sobre Tuberculose no estado de Pernambuco

**Quadro 1** – Clusters e grau epidemiológico da tuberculose em Pernambuco (2001-2024).

| CLUSTER   | GRAU EPIDEMIOLÓGICO             |
|-----------|---------------------------------|
| Cluster 1 | Baixa/Média incidência inicial. |
| Cluster 2 | Alta incidência persistente.    |
| Cluster 3 | Muito baixa incidência.         |
| Cluster 4 | Incidência intermediária.       |

**Fonte:** Os Autores.

A análise da distribuição etária, apresentada na Figura 4, revelou medianas variando entre 39 anos e 44 anos, aproximadamente. O cluster 2, responsável pela maior concentração de casos, apresentou a menor mediana (39 anos), indicando predominância da tuberculose em adultos economicamente ativos, o que é consistente com padrões epidemiológicos observados em populações urbanas e laboralmente expostas. Em contraste, o cluster 3 exibiu casos em faixas etárias mais elevadas (44 anos), enquanto os clusters 1 e 4 apresentaram medianas intermediárias (41 anos e 40 anos, respectivamente). A distribuição das medianas, associada à dispersão observada por meio da PCA, reforça a heterogeneidade epidemiológica entre os agrupamentos e destaca a idade como variável discriminante relevante no processo de clusterização.

**Figura 4** – Distribuição da idade média por cluster

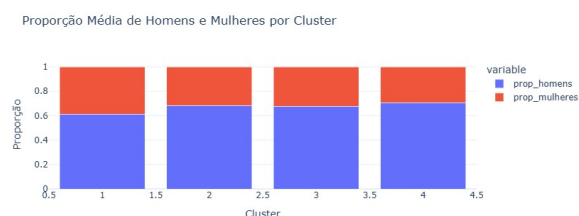


**Fonte:** Os Autores.

Quanto à análise da variável sexo, apresentado na Figura 5, observou-se predominância estatisticamente consistente do sexo masculino em todos os clusters, com proporções variando de 61% a 70% dos casos. Observou-se que a maior

prevalência masculina ocorreu nos clusters de maior incidência (2 e 4), indicando possível associação entre maior carga de adoecimento e a distribuição por sexo. Essa heterogeneidade percentual sugere diferenças estruturais entre os agrupamentos e reforça a relevância da variável sexo como fator discriminante no processo de clusterização. Além disso, a predominância masculina está alinhada com achados epidemiológicos prévios, que apontam maior vulnerabilidade de homens adultos ao adoecimento por tuberculose, tanto por fatores comportamentais quanto por barreiras de acesso aos serviços de saúde.

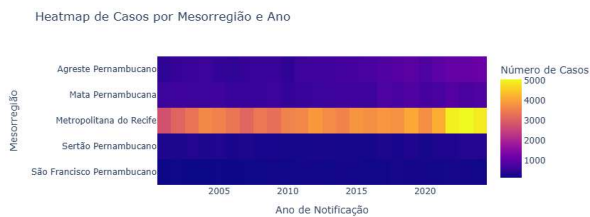
**Figura 5** – Distribuição proporcional de sexos na incidência de casos por cluster



**Fonte:** Os Autores.

O mapa de calor (*Heatmap*) de casos por mesorregião e ano evidenciou que a Região Metropolitana do Recife (RMR) concentrou os maiores valores em todo o período, enquanto as demais mesorregiões apresentaram incidência baixa a moderada. O comportamento observado na RMR sugere a presença de determinantes estruturais, como maior densidade populacional, intensificação da mobilidade urbana, desigualdades socioeconômicas e maior circulação de pessoas, fatores que historicamente favorecem a transmissão de doenças infecciosas. As outras mesorregiões caracterizam um cenário epidemiológico menos intenso e possivelmente associado à dispersão populacional, menor urbanização e menor pressão de transmissão. A heterogeneidade observada reforça a importância de estratégias de vigilância diferenciadas por território, considerando as especialidades epidemiológicas e sociais que moldam o comportamento espacial da doença.

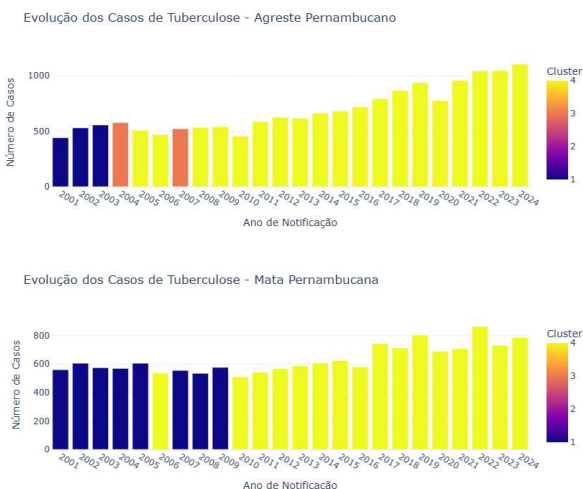
**Figura 3** – Heatmap dos casos de tuberculose por mesorregião e ano (2001-2024)



**Fonte:** Os Autores.

A evolução temporal mostrou que Agreste Pernambucano, Mata Pernambucana, Sertão e São Francisco Pernambucano iniciaram com perfis de baixa/média incidência (cluster 1), posteriormente oscilaram para o cluster 3, o que indica aumento pontual da intensidade epidemiológica em determinados anos. Ao final do período, essas regiões convergiram em perfil intermediário (cluster 4), sugerindo uma trajetória de crescimento moderado, porém sustentado. Já a RMR manteve-se consistentemente no cluster 2, ao longo de toda a série temporal, reafirmando-se como o principal epicentro da tuberculose em Pernambuco. O padrão de estabilidade elevada associado a valores sistematicamente superiores, ressalta a persistência do risco epidemiológico na região e reforça a necessidade de ações contínuas e localizadas de vigilância e controle. Essa análise da evolução temporal revelou um padrão de transição estatisticamente consistente entre os perfis epidemiológicos das mesorregiões.

**Figura 4** – Evolução dos casos de tuberculose por mesorregião e cluster (2001-2024)



Evolução dos Casos de Tuberculose - Sertão Pernambucano



Evolução dos Casos de Tuberculose - São Francisco Pernambucano



Evolução dos Casos de Tuberculose - Metropolitana do Recife



**Fonte:** Os Autores.

## 4.2 DISCUSSÃO

A RMR destacou-se como área crítica, com incidência elevada e persistente, expressando a necessidade de políticas intensivas adaptadas nessa região. As demais mesorregiões apresentaram trajetórias de transição: inicialmente em clusters de baixa/média incidência, migraram para perfis intermediários ao longo dos anos. Esse comportamento é condizente com estudos internacionais que identificam padrões diferenciados de progressão entre regiões de maior densidade populacional e áreas rurais de menor incidência [5].

A predominância masculina observada nos clusters com maior concentração (61-70%) está em concordância com evidências internacionais, que apontam maior vulnerabilidade dos homens à tuberculose devido a fatores sociais e comportamentais, como maior exposição ocupacional, prevalência de hábitos de risco e menor procura por serviços de saúde. Essa concentração reforça os resultados obtidos em estudos locais, que também identificaram maior prevalência da doença em homens jovens em idade produtiva [3].

Além disso, a diferença etária entre clusters, com casos concentrados em adultos jovens na RMR e em faixas etárias elevadas nas regiões de baixa incidência, reforça que estratégias de controle devem considerar a heterogeneidade demográfica das mesorregiões. Esse aspecto dialoga com análises internacionais que destacaram a importância de incorporar variáveis demográficas e socioeconômicas nos modelos de clusterização para melhor compreender a vulnerabilidade regional [12;6].

Os resultados evidenciam a necessidade de políticas adaptadas para cada mesorregião, moldadas ao perfil epidemiológico, e confirma o potencial da aplicação de técnicas de inteligência artificial, como a clusterização hierárquica, na análise territorial e epidemiológica tuberculose.

## **5 CONCLUSÕES**

Este estudo analisou os padrões epidemiológicos da tuberculose em Pernambuco entre 2001 e 2024, por meio de técnicas de clusterização hierárquica aplicadas às mesorregiões do estado com base nos indicadores disponibilizados pelo SINAN. Os resultados evidenciaram uma marcada heterogeneidade espacial e temporal da doença, refletida na distribuição desigual dos casos e na dinâmica diferenciada de transição entre clusters. A Região Metropolitana do Recife (RMR) apresentou alta estabilidade temporal, permanecendo continuamente alocada no cluster de maior incidência, sugerindo baixa variabilidade interanual e consolidando-se como epicentro da doença. Enquanto as demais mesorregiões apresentaram trajetórias não estacionárias, migrando de perfis de baixa/média incidência para níveis intermediários ao longo da série histórica, indicando um aumento gradual e estatisticamente consistente da carga epidemiológica.

A análise demográfica reforçou os achados, revelando predominância masculina estatisticamente consistente (61-70%) em todos os clusters, especialmente nos de alta incidência. Em acordo com as evidências internacionais que apontam maior vulnerabilidade dos homens por fatores sociais, comportamentais e relacionados ao acesso aos serviços de saúde. Contribuindo para taxas mais elevadas de transmissão e diagnóstico tardio. As diferenças etárias entre clusters – com

casos concentrados em adultos jovens na RMR e em faixas mais elevadas nas regiões de baixa incidência – reforçam a necessidade de políticas públicas segmentadas ao perfil epidemiológico e demográfico de cada região.

Os resultados obtidos conversam com estudos anteriores que aplicaram técnicas de mineração de dados em epidemiologia, como análises de tuberculose em Recife [3] e investigações internacionais com K-Means e K-Medoids [5;6]. No entanto, o presente trabalho diferencia-se ao empregar clusterização hierárquica, que oferece vantagens como visualização detalhada das relações inter-regionais por meio do dendrograma, possibilitando interpretações mais ricas e decisões mais fundamentadas para definição do número ótimo de clusters. Além disso, robustez na identificação de padrões multiescala, essencial para fenômenos epidemiológicos que variam simultaneamente no tempo e no espaço. É uma melhor interpretação da progressão epidemiológica, já que diferentes “alturas” do dendrograma revelam não apenas clusters finais, mas também agrupamentos intermediários que ajudam a compreender trajetórias evolutivas das regiões. Assim, o uso de clusterização hierárquica não apenas permitiu identificar agrupamentos coerentes e estatisticamente consistentes, como também forneceu subsídios interpretativos relevantes para a compreensão da dinâmica epidemiológica da tuberculose em Pernambuco, fortalecendo o potencial da técnica como ferramenta de apoio à vigilância e ao planejamento em saúde pública.

Conclui-se, portanto, que a aplicação de métodos de inteligência artificial, como a clusterização hierárquica, representa uma ferramenta promissora para a análise territorial e epidemiológica da tuberculose. Essa técnica demonstrou robustez em relação a métodos particionais, como K-Means e K-Medoids, ao permitir a identificação de agrupamentos em diferentes níveis de agregação e fornecer subsídios interpretativos relevantes para a compreensão da dinâmica epidemiológica. Pesquisas futuras devem integrar variáveis socioeconômicas e ambientais, ampliando a capacidade explicativa dos modelos e fortalecendo a aplicabilidade em políticas públicas. Ademais, recomenda-se, ainda, a padronização da incidência por 1.000 habitantes nas mesorregiões e

a inclusão de novos indicadores epidemiológicos e demográficos, ampliando a robustez das análises e contribuindo para o planejamento de políticas públicas mais eficazes e direcionadas.

## REFERÊNCIAS

[1] BRASIL. Ministério da Saúde. **Tuberculose**. Brasília: Ministério da Saúde, 2025. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/t/tuberculose>. Acesso em: 30 nov. 2025.

[2] BRASIL. Ministério da Saúde. **Apresentação: dados epidemiológicos da tuberculose no Brasil**. Brasília: Ministério da Saúde, 2024. Disponível em: <https://www.gov.br/saude/pt-br/centrais-de-conteudo/apresentacoes/2024/apresentacao-dados-epidemiologicos-da-tuberculose-no-brasil>. Acesso em: 30 nov. 2025.

[3] FARIAS, Alessandra Nunes et al. Epidemiologia da tuberculose e sua distribuição espacial na Região Metropolitana de Recife entre 2019 e 2020. **Brazilian Journal of Infectious Diseases**, 2021. Disponível em: <https://doi.org/10.1016/j.bjid.2021.101980>. Acesso em: 30 nov. 2025.

[4] ROSSINI, Angelo et al. Unsupervised K-means analysis of tuberculosis data in Brazil: identifying high prevalence states and temporal trends. **Procedia Computer Science**, v. 239, p. 1839-1846, 2024. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050924016090>. Acesso em: 11 set. 2025.

[5] AZIZAH, Ayu; ILHAM, Amil Ahmad; SYAFARUDDIN. Spatial analysis of the spread of tuberculosis cases based on socio-economic factors using distance-based algorithm. **IECELTICS: International Conference on Electrical Engineering and Informatics**, IEEE, 2022.

[6] AFIFAH, Sarah Ummul et al. Spatial-temporal visualization of tuberculosis vulnerability in East Java, Indonesia, using the K-medoids clustering algorithm. **IES: International Electronics Symposium**, Surabaya, 2022.

[7] ALVES, Geovanne Oliveira et al. Estudo comparativo de técnicas de previsão para casos de dengue. **Revista de Engenharia e Pesquisa**

**Aplicada**, Recife, v. 6, n. 3, p. 12-20, 2021. DOI: <https://doi.org/10.25286/rep.v6i3.168312>.

[8] ROCHA, Allana et al. Aplicação de clustering para segmentação de clientes na base de dados da Justa. **Revista de Engenharia e Pesquisa Aplicada**, Recife, v. 7, n. 3, p. 39-53, 2021. DOI: <https://doi.org/10.25286/rep.v7i3.2458>.

[9] BORGES, Regis et al. Aplicando clusterização para identificação de grupos em base de dados do questionário CUDIT. **Revista de Engenharia e Pesquisa Aplicada**, Recife, v. 7, n. 2, p. 42-47, 2022. DOI: <https://doi.org/10.25286/rep.v7i2.2216>.

[10] BRASIL. Ministério da Saúde. Departamento de Informática do SUS – DATASUS. **Transferência de Arquivos**. Brasília: Ministério da Saúde, 2025. Disponível em: <https://datasus.saude.gov.br/transferencia-de-arquivos/>. Acesso em: 18 jul. 2025.

[11] SILVA, Marcus Tolentino; GALVÃO, Taís Freire. Incidência de tuberculose no Brasil: análise de série temporal entre 2001 e 2021 e projeção até 2030. **Revista Brasileira de Epidemiologia**, São Paulo, v. 27, e202500, 2024. Disponível em: <https://www.scielo.br/rbepid>. Acesso em: 30 nov. 2025.

[12] PATHAK, Abhijit et al. Application of machine learning K-means clustering and linear regression in determining the risk level of pulmonary tuberculosis. **COMPAS: IEEE INTERNATIONAL CONFERENCE ON COMPUTING, APPLICATIONS AND SYSTEMS**, 2024, Bangladesh.