Universidade de São Paulo
Instituto de Matemática e Estatística

Centro de Estatística Aplicada

Relatório de Análise Estatística

RAE-CEA–25P06

RELATÓRIO DE ANÁLISE ESTATÍSTICA SOBRE O PROJETO:

Ocorrência da esporotricose felina no município de São Paulo e a desratização como medida ambiental de controle

Matheus de Castro Siniscarchio
Murilo Fernandes da Costa
Rafael Bassi Stern

São Paulo, junho de 2025

CENTRO DE ESTATÍSTICA APLICADA - CEA – USP

TÍTULO: Relatório de Análise Estatística sobre o Projeto: “Ocorrência da esporotricose felina no município de São Paulo e a desratização como medida ambiental de controle”.

PESQUISADORA: Camila Baltazar

ORIENTADOR: Prof. Dr. Nilson Roberti Benites

INSTITUIÇÃO: Faculdade de Medicina Veterinária e Zootecnia (FMVZ-USP)

FINALIDADE DO PROJETO: Doutorado

RESPONSÁVEIS PELA ANÁLISE: Matheus de Castro Siniscarchio
Murilo Fernandes da Costa
Rafael Bassi Stern

REFERÊNCIA DESTE TRABALHO: SINISCARCHIO, M. C.; COSTA, M. F.; STERN, R. B. Relatório de análise estatística sobre o projeto: “Ocorrência da esporotricose felina no município de São Paulo e a desratização como medida ambiental de controle”. São Paulo, IME-USP, 2025. (RAE–CEA-25P06)

FICHA TÉCNICA

REFERÊNCIAS BIBLIOGRÁFICAS:

DIAS, R.A. (2013). **Os donos do pedaço: caracterização das populações de cães e gatos domiciliados no município de São Paulo**. Universidade de São Paulo. Faculdade de Medicina Veterinária e Zootecnia. DOI: 10.11606/9788567421018 Disponível em: www.livrosabertos.abcd.usp.br/portaldelivrosUSP/catalog/book/322 . Acesso em 1 junho. 2025.

Secretaria Municipal da Saúde de São Paulo. (2017). Cães e gatos no Município de São Paulo: Imunização, esterilização e convivência com humanos. **Boletim ISA Capital 2015**, 8. São Paulo: Coordenação de Epidemiologia e Informação - CEInfo.

GOOGLE. *Geocoding API*. Google Developers. Disponível em: https://developers.google.com/maps/documentation/geocoding. Acesso em: 12 jul. 2025.

BOEING, Geoff. *OSMnx: Python for street networks*. Documentação da biblioteca OSMnx. Disponível em: https://osmnx.readthedocs.io. Acesso em: 12 jul. 2025.

ESTER, Martin; KRIEGEL, Hans-Peter; SANDER, Jörg; XU, Xiaowei. *A density-based algorithm for discovering clusters in large spatial databases with noise*. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining – KDD-96. Portland, OR: AAAI Press, 1996. p. 226–231.

PROGRAMAS COMPUTACIONAIS UTILIZADOS:

Microsoft Word for Windows (versão 2024)
Microsoft Excel for Windows (versão 2024)
Jupyter Notebook (versão 7.0.8)
Python (versão 3.11.7)

TÉCNICAS ESTATÍSTICAS UTILIZADAS:

03:020 – Análise Descritiva Multidimensional

03:990 – Outros

ÁREA DE APLICAÇÃO

14:010 – Ciências Físicas e Geoestatística

Resumo

O relatório trata da ocorrência de esporotricose felina no município de São Paulo, uma micose zoonótica causada pelo fungo Sporothrix, que afeta gatos e humanos. O aumento expressivo de casos da doença nos últimos anos levou à necessidade de identificar padrões espaciais e temporais da sua disseminação. Utilizando dados da Secretaria Municipal de Saúde entre 2011 e 2024, o estudo procurou mapear os focos da doença com base na geolocalização dos registros e dados socioeconômicos distritais, a fim de embasar políticas públicas de controle, como a desratização.

A análise se concentrou em três bases de dados: registros de casos, estimativas populacionais de felinos e o “Mapa da Desigualdade” de 2023, que traz 54 variáveis socioeconômicas por distrito. A partir desses dados, foi possível identificar que os distritos mais afetados — como Cangaíba, Sapopemba e São Mateus — também apresentam altos índices de desigualdade. Houve uma correlação negativa significativa entre a ocorrência da doença e o índice de desigualdade (Spearman = -0,665), indicando maior vulnerabilidade em regiões periféricas.

Para identificar focos ativos da doença, foi utilizada a técnica de agrupamento DBSCAN, que permite detectar aglomerações de casos independentemente de limites administrativos. Foram identificados 28 focos, principalmente nas zonas leste e norte da cidade. Isso reforça a importância de ações específicas nessas áreas, considerando a complexidade das rotas de disseminação da doença em grandes centros urbanos.

Além disso, foram testados modelos preditivos para estimar o risco de novos casos por distrito. O modelo final considerou como variáveis explicativas o índice de desigualdade e o histórico de casos em até quatro trimestres anteriores. Embora os modelos não tenham alcançado precisão absoluta, atingiram bons níveis de acerto — até 87% — especialmente na previsão de distritos com risco alto, que são os mais relevantes para a alocação de recursos.

Conclui-se que a integração entre dados espaciais, estatísticas sociais e epidemiológicas fornece uma base sólida para políticas públicas mais eficazes. O estudo evidencia a importância de modelos preditivos que considerem desigualdades estruturais e padrões de disseminação para o controle da esporotricose felina, contribuindo para uma abordagem mais precisa e eficiente na saúde pública urbana.

Sumário

1. Introdução 7

2. Objetivo(s) 7

3. Descrição do estudo 8

4. Descrição das variáveis 8

4.1 Amostra dos registros 9

4.2 Mapa da desigualdade 9

4.3 População de felinos 13

5. Análise descritiva 13

5.1 Geolocalização dos dados 13

5.2 Divisão Regional de São Paulo 14

5.3 Quantidade de casos totais por Distrito 15

Quantidade de gatos por distrito 16

Análise da proporção de ocorrência de doenças em felinos por distrito 16

6. Análise inferencial 17

7. Conclusões 17

APÊNDICE A 18

APÊNDICE B 19

1. Introdução

A esporotricose felina é uma micose zoonótica emergente, causada por fungos do gênero Sporothrix, que afeta tanto animais quanto humanos, manifestando-se principalmente como lesões cutâneas, linfáticas ou sistêmicas. No Brasil, a doença ganhou destaque devido a surtos recentes, especialmente no município de São Paulo, onde foi registrado um crescimento alarmante de casos nos últimos anos. Esse cenário evidencia a necessidade de estratégias eficazes de vigilância e controle para mitigar a disseminação da doença.

O projeto busca investigar a dinâmica espaço-temporal da disseminação da esporotricose felina por meio de georreferenciamento, utilizando as definições geográficas do IBGE para delimitação de áreas de estudo. A pesquisa visa mapear a expansão da doença com base em dados históricos (2011–2024) da Secretaria Municipal de Saúde, identificando padrões de disseminação e classificando regiões por nível de risco.

A abordagem integra técnicas de análise espacial e estatística, alinhadas às definições geográficas do IBGE para garantir precisão na delimitação territorial e na interpretação dos dados.

Os resultados esperados incluem a criação de um modelo preditivo para a disseminação da esporotricose. Esses achados poderão embasar políticas públicas direcionadas, otimizando recursos e fortalecendo ações de vigilância em saúde única.

O projeto destaca-se por sua aplicabilidade prática, contribuindo para o controle desta zoonose em ambientes urbanos.

2. Objetivo

O projeto tem como objetivo a criação de um modelo que quantifique o risco de expansão da doença de acordo com a localidade. Essa quantificação servirá para que se entenda melhor onde estão surgindo novos focos da doença para que seja possível alocar recursos direcionados para a prevenção da doença nesses locais.

Dada a limitação de recursos para essa prevenção, é de interesse da pesquisadora que essa quantificação se dê na menor divisão regional possível, pois isso permitiria uma ação mais direcionada e precisa.

Outra necessidade é integrar a esse modelo o nível socioeconômico de cada região estudada, visto que existe uma correlação entre a capacidade de expansão da doença e o acesso a recursos e a tratamento de cada região. Dessa forma, um modelo que leve em consideração essas diferenças é muito mais poderoso em sua habilidade de previsão.

3. Descrição do estudo

Os dados foram coletados no período de 2011 até 2024, e referem-se a registros de ocorrências de esporotricose dentro de cada UVI do município de São Paulo. As UVI's são unidades descentralizadas que atuam em vigilância ambiental, sanitária e epidemiológica, coordenadas pela COVISA (Coordenação de Vigilância em Saúde) da Prefeitura de São Paulo.

A esporotricose se manifesta no gato através de pequenas feridas em seu corpo que muitas vezes não são identificadas imediatamente como uma doença. Por esse motivo, existe um certo atraso no tempo entre o momento em que o animal foi contaminado e o momento em que o animal foi diagnosticado e registrado. Na base de dados encontra-se a data de registro da doença, e a data estimada de início da contaminação, no entanto essa segunda não está presente em todos os dados e é uma estimação dada pelo dono do animal, por isso para esse projeto será utilizada a data de registro da doença.

Os dados utilizados para classificar socioeconomicamente cada região foram obtidos de um estudo de 2023 chamado "Mapa da desigualdade". Esse estudo criou um índice que classifica cada distrito do município de São Paulo de acordo com a desigualdade econômica do distrito. Para isso, foram utilizadas 54 variáveis de vários temas diferentes que seriam indicadoras da desigualdade local.

4. Descrição das variáveis

O projeto trabalha com 3 bancos de dados distintos, sendo que o primeiro contém os casos registrados entre 2011 e 2024, o segundo contém as informações do estudo do "Mapa da desigualdade" e o terceiro as estimativas para a população de felinos.

4.1 Casos registrados

Ficha nº - Número de identificação do registro

Ano - Ano de ocorrência do caso

Data notificação - Data em que foi realizada a notificação (dd/mm/aaaa)

Logradouro - Nome da rua/avenida onde ocorreu o caso

Número - Número do endereço do local

CEP - Código de Endereçamento Postal do local

D.A. - Distrito administrativo onde ocorreu o caso

Início lesão - Data em que surgiram os primeiros sinais da lesão (dd/mm/aaaa)

4.2 Mapa da desigualdade

O estudo “Mapa da desigualdade” possui um conjunto de informações sobre todos os distritos de São Paulo que são indicadores da desigualdade de cada local. São no total 54 variáveis divididas nos temas: “População”, “Meio ambiente”, “Mobilidade”, “Direitos humanos”, “Habitação”, “Saúde” e “Educação”.

População total - População total, por distrito

População preta e parda - Proporção (%) da população preta e parda, por distrito

População feminina - Proporção (%) da população feminina, por distrito

População jovem - Proporção (%) de população de 0 a 29 anos, por distrito

População Infantil - Proporção (%) de população de 0 a 6 anos, por distrito

População em situação de rua - População total em situação de rua, por distrito

Emissão de poluentes atmosféricos por área (kg/km²/dia) - Emissão de material particulado associada às viagens de ônibus, gerado por combustão e por desgaste de pneus, freios e pistas (kg), sobre área do distrito paulistano onde a emissão ocorreu (km²), para um dia útil típico do ano de 2021.

Cobertura vegetal - Percentual de áreas com cobertura vegetal (%)

Pontos de entrega voluntária (PEV) - Número de pontos de entrega voluntária (PEV)

Número de áreas contaminadas - Número de áreas contaminadas

Mortes no trânsito - Coeficiente de mortes em acidentes de trânsito para cada cem mil habitantes, por distrito

Velocidade média ônibus - Velocidade média (km/h) dos ônibus, por distrito

Tempo médio de deslocamento por transporte público - Tempo médio (em minutos) de deslocamento por transporte público (pico da manhã), por distrito

Violência contra a mulher - Coeficiente de mulheres vítimas de violência (todas as categorias) para cada dez mil mulheres residentes de 20 a 59 anos, por distrito

Violência LGBTQIAP+ - Coeficiente de pessoas vítimas de violência homofóbica e transfóbica para cada cem mil habitantes, por distrito

Violência racial - Coeficiente de pessoas vítimas de violência de racismo e injúria racial para cada dez mil habitantes, por distrito

Famílias em atendimento habitacional provisório - Famílias em atendimento habitacional provisório por situação de risco e emergência

Favelas - Proporção (%) estimada de domicílios em favelas em relação ao total de domicílios, por distrito

Gravidez na adolescência - Proporção (%) de nascidos vivos de parturientes com menos de 20 anos em relação ao total de nascidos vivos

Mortalidade materna – Razão de mortalidade materna, Média trienal (2020 a 2022) do total de óbitos por causas maternas dividido pela média trienal (2019 a 2021) do total de nascidos vivos multiplicado por 100.000.

Mortalidade infantil - Coeficiente de mortalidade infantil, para cada mil crianças nascidas vivas de mães residentes no distrito

Idade média ao morrer - Média de idade (em anos) das pessoas que morreram (de acordo com o local de residência), por distrito

Tempo médio para consultas na atenção básica - Tempo médio (em dias) de espera para consultas na atenção primária

Mortalidade por Covid-19 - Proporção (%) de óbitos por covid-19 em relação ao total de óbitos

Tempo de atendimento para vaga em creche - em dias

Matrículas no ensino básico em escolas públicas - Proporção (%) de matrículas no Ensino Básico em escolas públicas e conveniadas em relação ao total de matrículas, por distrito

Distorção idade-série no ensino fundamental da rede municipal - Taxa de distorção idade-série (Ensino Fundamental da Rede Municipal). Total de alunos do Ensino

Fundamental da rede municipal, matriculados com idade acima da recomendada para a série (idade recomendada +2) ÷ Total de matrículas no Ensino Fundamental da rede municipal x 100

Abandono escolar no ensino fundamental da rede municipal - Proporção (%) de alunos que abandonaram a escola no Ensino Fundamental da rede municipal

Ideb (Escolas públicas - anos iniciais) - Nota média do Ideb para as escolas públicas do Ensino Fundamental (anos iniciais)

Ideb (Escolas públicas - anos finais) - Nota média do Ideb para as escolas públicas do Ensino Fundamental (anos finais)

Adequação da formação docente - Proporção (%) média de docentes do Ensino Fundamental com formação inadequada a disciplina que lecionam

Esforço docente - Proporção (%) média de docentes do Ensino Fundamental com alto grau de esforço docente

Centros culturais, casas e espaços de cultura - Proporção (%) de centros culturais, espaços e casas de cultura (municipais), para cada dez mil habitantes, por distrito

Equipamentos públicos de cultura - Proporção (%) de equipamentos públicos de cultura (municipais), para cada cem mil habitantes, por distrito

Cinemas - Proporção (%) de salas de cinema (municipais), para dez mil habitantes, por distrito

Espaços culturais independentes - Número de espaços culturais independentes, para cada cem mil habitantes, por distrito

Equipamentos públicos de esporte - Número de equipamentos esportivos públicos (municipais e estaduais) de esporte para cada dez mil habitantes, por distrito.

Quadras esportivas nas escolas públicas - Proporção (%) de escolas públicas com quadra esportiva, em relação ao total de escolas públicas

Oferta de emprego formal - Taxa de oferta de emprego formal, por dez habitantes participantes da população em idade ativa (PIA), por distrito

Remuneração média mensal do emprego formal - Remuneração média mensal (em R$) do emprego formal, por distrito

Desigualdade salarial (emprego formal) - Desigualdade de salário por sexo (salário de mulheres / salário de homens)

Acesso a transporte de massa - Proporção (%) da população que reside em um raio de até 1 km de estações de sistemas de transporte público de alta capacidade (trem, metrô e monotrilho), por Zona OD (origem-destino) e por distrito

Acesso a infraestrutura cicloviária - Proporção (%) da população que reside em um raio de até 300 metros de distância de infraestruturas cicloviárias (ciclovias e ciclofaixas), por Zona OD (origem-destino) e por distrito

Acesso internet móvel (por área/km²) - Distribuição de antenas (Estações Radio-base - ERBs), por área (km²) dos distritos

Acesso internet móvel (População distrito) - Distribuição de antenas de internet móvel a cada dez mil habitantes, por distrito

Feminicídio - Coeficiente de mulheres vítimas de feminicídio, para cada dez mil mulheres residentes de 20 a 59 anos, por distrito

Homicídios - Coeficiente mortalidade por homicídio e intervenção legal para cada cem mil pessoas residentes, por distrito

Homicídios de jovens - Coeficiente estimado mortalidade de jovens por homicídio e intervenção legal para cada cem mil pessoas residentes de 15 a 29 anos, por distrito

Agressões por intervenção policial - Coeficiente estimado de violência em intervenção legal registradas nas unidades de saúde para cada 100 mil habitantes, por distrito.

Mortes por intervenção policial - Coeficiente estimado de casos registrados em boletins de ocorrência na categoria mortes decorrentes de intervenção policial (MDIP) para cada 100 mil habitantes, por distrito.

Deslocamentos médio para denúncias de violência contra mulher - Deslocamento médio (km) de mulheres vítimas de violência (todas as categorias) por distrito

Coeficiente de desigualdade - Valor calculado através das 51 variáveis citadas acima para ordenar os distritos como mais ou menos desiguais. Um valor maior desse coeficiente significa que o distrito é menos desigual.

Esses dados são muito relevantes pois possuem informações precisas e recentes sobre cada distrito. É provável que nem todas sejam utilizadas no modelo final, porém muitas delas podem ter uma correlação alta com o crescimento da doença naquela região e assim ajudar na capacidade preditiva do modelo.

4.3 População de felinos

A estimativa da população de felinos foi feita através de dois estudos. O primeiro é uma publicação de 2009 “Os donos do pedaço” (DIAS, 2013), sendo que nela se encontram dados de 3 variáveis de interesse para esse projeto:

População - População humana para cada distrito em 2009.

População felina - População estimada de felinos para cada distrito em 2009.

Razão homem:gato em 2009- Razão estimada entre a população humana e felina para cada distrito em 2009.

O segundo é a pesquisa “ISA Capital 2015” da Secretaria Municipal da Saúde de São Paulo (2017), que possui dados de apenas uma variável:

Razão homem:gato em 2015 - Razão estimada entre a população humana e felina para cada região do município de São Paulo, em 2015.

O estudo de 2009 é mais completo e possui informações sobre cada distrito, porém o de 2015 é mais recente e talvez seja o mais adequado por estar mais atualizado. Para essa análise descritiva, a população de felinos foi estimada utilizando a razão de cada região em 2015 e a população humana de cada distrito em 2022, porém considerando a variabilidade dentro de cada região para a criação do modelo haverá uma tentativa de melhor estimar essa população ponderando essa variação.

5. Análise descritiva

5.1 Geolocalização dos dados

A geolocalização dos endereços de ocorrência é essencial para permitir a análise espacial e o agrupamento dos casos segundo as divisões geográficas do IBGE. O processo foi realizado em duas APIs. API é uma Interface de Programação de Aplicações, que tem como objetivo nos permitir “conversar com um banco de dados” e extrair informações.

1. **OpenStreetMap (OSM)**

Utilizou-se inicialmente a API Nominatim do OSM, de código aberto.

Identificaram-se limitações devido à alta sensibilidade a erros de digitação ou inconsistências nos endereços (ex.: abreviações, caracteres ausentes).

A falha na geocodificação ocorria mesmo para pequenas divergências, resultando em baixa completude.

2. Google Maps Platform:

Adotou-se a API de Geocodificação do Google, que inclui recursos de fuzzy matching para corrigir erros em endereços.

Essa abordagem garante alta taxa de sucesso na geolocalização, mesmo para registros com imperfeições.

Apesar de ser uma solução paga, utilizou-se a licença gratuita de três meses para projetos não comerciais, viabilizando o processamento dos dados.

5.2 Divisão regional de São Paulo

O IBGE classifica o território brasileiro em diversas divisões geográficas para fins estatísticos e administrativos. Entre essas classificações, os distritos e setores censitários são fundamentais para a organização dos censos demográficos e serão utilizados como unidades de agrupamento para os casos de ocorrência dos gatos.

- Distritos: São subdivisões dos municípios, podendo conter subdistritos em algumas cidades. Cada distrito tem um centro administrativo e pode englobar áreas urbanas e rurais.
- Setores censitários: São as menores unidades territoriais utilizadas pelo IBGE para coleta de dados censitários. Cada setor censitário corresponde a uma área específica dentro de um distrito, delimitada para facilitar a pesquisa de campo.

O código do setor censitário do IBGE é um geocódigo único que identifica cada setor censitário dentro do território nacional. Ele é composto por uma sequência numérica que reflete a hierarquia territorial, incluindo:

1. Código da Unidade da Federação (UF) – Representa o estado onde o setor está localizado.
2. Código do Município – Identifica o município dentro da UF.
3. Código do Distrito – Define o distrito dentro do município.

4. Código do Subdistrito (se houver)
5. Código do Setor Censitário – É o identificador específico do setor dentro do distrito.

Cada setor censitário está integralmente contido dentro de um distrito, respeitando a divisão político-administrativa do Brasil. Essa estrutura permite que os dados coletados sejam organizados de forma precisa para análises estatísticas e planejamento urbano.

Além disso, a escolha do setor censitário se deu por conta da facilidade da transição dentro do setor, já que ele foi feito para que recenseadores pudessem coletar informações para o censo.

5.3 Quantidade de casos totais por ano

A Figura B.1 apresenta a quantidade de ocorrências de esporotricose agregada por ano, evidenciando um crescimento significativo ao longo do período de 2011 a 2024. Nos primeiros anos, os números eram relativamente baixos, mas a partir de 2016 observa-se um aumento expressivo, que se intensifica após 2018

5.4 Quantidade de casos totais por Distrito

Primeiramente, consideramos um agregado de todos os casos por distrito, sem considerar o tempo. Dessa forma, conseguimos quantificar onde estão as regiões com a maior quantidade de casos. Conforme se observa na Figura B.2:

- Os distritos com os maiores números de casos estão espalhados por diferentes regiões da cidade, indicando que não há uma concentração exclusiva em uma única área.
- Há uma predominância de casos em regiões periféricas, o que pode indicar fatores socioeconômicos influenciando a distribuição.
- Os distritos nomeados se encontram na faixa de 396 a 1277 casos, o que os coloca entre os mais afetados.

A Figura B.3 mostra um exemplo de abertura dos casos em setores censitários para evidenciar que os casos não são distribuídos uniformemente dentro de um distrito.

Por conta disso, é interessante analisar a nível do setor censitário; porém, por conta da alta quantidade de setores no município de São Paulo (27.301) não é possível analisar de forma descritiva individual.

5.5 Quantidade de gatos por distrito

A quantidade de gatos foi estimada pela pesquisadora a partir da pesquisa de 2015, onde se estimou a proporção de gatos por pessoa. Ela replicou esses números para a população do censo de 2022.

Dessa forma, conseguimos entender melhor como está ocorrendo a doença com base no número de gatos por pessoa.

5.6 Análise da proporção de ocorrência de doenças em felinos por distrito

A Figura B.4 apresenta a proporção de casos de doença em felinos por distrito de São Paulo, ou seja, a quantidade de ocorrências da doença agrupados desde 2018 sobre a quantidade de gatos estimada. A escala de cores varia de **0.00 a 0.07**, com tons mais escuros indicando distritos onde há uma maior incidência proporcional da doença em relação à população total de gatos.

Os distritos destacados – **Cangaíba, Vila Maria, Aricanduva, São Mateus e Sapopemba** – possuem os maiores índices de ocorrência, sugerindo padrões relevantes.

5.7 Correlação da quantidade de ocorrência da doença por distrito com índices de desigualdade

A Figura B.5 analisa a relação entre a quantidade de casos da doença e o coeficiente de desigualdade (quanto mais baixa a pontuação, maior a desigualdade). O valor do coeficiente de correlação de Spearman (-0,665) revela uma **relação monotônica negativa**, sugerindo que a tendência inversa (mais desigualdade → mais doença) é consistente.

6. Análise inferencial

6.1 Identificação de focos da doença

Devido à dificuldade de localizar focos da doença, causada tanto pela ampla dispersão geográfica dos casos quanto pelo elevado número de ocorrências, foi aplicada uma técnica de agrupamento conhecida como **DBSCAN** (*Density-Based Spatial Clustering of Applications with Noise*). Essa técnica permite identificar focos de forma

objetiva, combinando três critérios principais: **a distância máxima permitida entre os casos**, **o número mínimo de casos necessário para formar um foco** e **o período histórico (em meses) considerado para o agrupamento**.

De maneira simplificada, o método agrupa os casos que estão próximos entre si. Quando o número mínimo de casos dentro da distância estabelecida é atingido, considera-se que há um foco de doença. Caso contrário, os casos são classificados como **outliers**, ou seja, não pertencem a nenhum foco.

Um aspecto importante é que o DBSCAN **não utiliza um "raio fixo" ao redor de cada caso**, mas sim trabalha com a relação de proximidade entre os casos de forma encadeada. Isso permite identificar focos com formatos irregulares ou não óbvios, como agrupamentos alongados, dispersos ou com geometrias incomuns, que outras técnicas baseadas apenas em distância direta poderiam não detectar.

Outra vantagem dessa abordagem é que ela dispensa a necessidade de agrupar os casos previamente por qualquer camada geográfica administrativa, como bairros, municípios ou regiões. O algoritmo trabalha diretamente com as coordenadas dos casos, permitindo a detecção de focos que podem atravessar esses limites administrativos.

Para exemplificar, a técnica foi aplicada para os casos da doença com um histórico dos últimos 6 meses, com uma distância entre observações de 1 km e a quantidade mínima de casos como 10. Na Figura B.6 é possível observar o resultado, 28 focos da doença foram encontrados em toda a cidade, com os maiores focos na zona leste e zona norte.

6.2 Previsão

Com o objetivo de investigar variáveis que possam contribuir para a predição da ocorrência de casos da doença, foram aplicadas técnicas estatísticas voltadas à previsão do número de casos. Para isso, utilizaram-se variáveis do Censo Demográfico de 2022, disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), juntamente com a proporção de humanos e gatos, estimada a partir de dados da Secretaria Municipal da Saúde de São Paulo (2017). Essa proporção foi utilizada para estimar a quantidade de gatos por setor censitário.

Além disso, para incorporar a dimensão espacial à modelagem, foi criada uma covariável representando a quantidade de gatos com a doença nos setores censitários vizinhos. Para a construção dessa variável espacial, elaborou-se um grafo a partir do mapa dos setores censitários do município de São Paulo, permitindo a identificação dos setores adjacentes. Dessa forma, a modelagem pôde capturar não apenas o efeito das características locais de cada setor, mas também a influência espacial e os padrões de desigualdade na distribuição dos casos.

Inicialmente, a tentativa de previsão seria feita agrupando quantidade de registros de casos por setor censitário a cada mês. No entanto, esse tipo de previsão resultou em um problema, visto que, como setor censitário é uma divisão geográfica muito pequena, o número de casos por mês em cada setor censitário era extremamente baixo. Mais de 99% das observações eram de 0 casos naquele setor censitário naquele mês, o que torna a previsão muito menos precisa, pois, a variabilidade é muito baixa. Os modelos iriam subestimar o número de casos pois se eles predisserem que sempre terá um número próximo de 0 de casos eles acertariam mais de 99% das vezes.

Por esse motivo, se optou por diminuir o agrupamento para um escopo maior. O agrupamento escolhido foi casos por distrito por trimestre. Essa divisão trará uma visão menos precisa de onde os casos irão surgir, já que os distritos são divisões geográficas bem maiores. Porém, acreditava-se que em conjunto com a identificação de focos essa previsão poderia trazer um entendimento melhor de onde é mais necessário alocar recursos.

Foram testados dois modelos diferentes para previsão. O primeiro é um modelo GLM (modelo linear generalizado) de quase-verossimilhança utilizando distribuição Poisson. A distribuição Poisson foi utilizada pois ela costuma modelar bem dados de contagem, como é o caso estudado, e a quase-verossimilhança ajuda a conter problema de superdispersão (variância muito maior que a média) que é algo que acontecia com esses dados devido a correlação entre as variáveis resposta.

No segundo modelo foi tentada a criação de um modelo SARIMA de séries temporais para a série de casos de cada distrito. Esse modelo, no entanto, não ia ter uma eficiência alta pois seria necessária a modelagem da série de cada distrito trimestralmente por um estatístico, porque a ordem de cada série pode mudar com o

aparecimento de novos casos e a automação dessa modelagem não é tão robusta e confiável. Devido à falta de aplicabilidade optou-se pelo primeiro modelo.

Considerando que o acerto preciso do número de casos não era de tanto interesse da pesquisadora, mas sim de saber se haveria ou não muitos casos naquela região, optou-se por uma transformação da variável numérica quantidade de casos em uma variável categórica. Isso foi feito criando duas maneiras distintas de classificar o número de casos. Sendo N o número de casos em um certo distrito em um certo trimestre, temos:

Primeira classificação:

Risco Baixo: $N < 10$

Risco Moderado: $10 \leq N < 30$

Risco Alto: $N \geq 30$

Segunda classificação:

Risco Baixo: $N < 15$

Risco Alto: $N \geq 15$

Os limiares dessas classificações foram escolhidos tentando maximizar a porcentagem de acertos de predição e a opção de usar uma outra ou até de mudar os limites dessas classificações pode ser feita pela pesquisadora de acordo com os resultados apresentados aqui e de acordo com o que ela achar que possui mais aplicabilidade.

Existe um problema em modelagem estatística chamado *overfitting*, que é quando o modelo se ajusta bem demais para um conjunto de dados específico, porém qualquer coisa fora daquilo ele terá uma predição ruim. Para evitar esse tipo de problema se separa os dados em conjuntos de treino e de teste. Dessa forma se ajusta o modelo no conjunto de treino e se usa o resto para o conjunto de teste evitando que eles se ajustem apenas a um conjunto específico e verificando se o modelo captura bem a variabilidade dos dados. A validação cruzada é um dos métodos de separação em treino e teste que faz isso repetidas vezes para diferentes separações para estimar o erro de previsão.

Na previsão do modelo, a variável resposta, que é o que estamos querendo prever, se refere a quantidade de casos da doença em um certo distrito no próximo trimestre. Havia várias variáveis que serviam como candidatas a variáveis explicativas, que seriam as utilizadas para prever a variável resposta. As variáveis temporais, se referem à

quantidade de casos da doença no mesmo distrito em trimestres anteriores, as variáveis do estudo do “Mapa da Desigualdade” e a variável referente à quantidade de casos da doença nos distritos vizinhos ao da variável resposta no último trimestre.

Em relação às variáveis do “Mapa da Desigualdade” havia uma escolha a ser feita entre usar um subconjunto das variáveis, ou usar o índice de desigualdade que o estudo calculava com base nas outras variáveis. Utilizando métodos de validação cruzada tentando minimizar o erro quadrático médio (erro de previsão), o modelo que havia menor erro era o que incluía todas as covariáveis do estudo.

Depois dessa validação, foram feitas previsões para os trimestres de 2022 até 2024 utilizando os trimestres anteriores ao trimestre previsto. As Tabelas A.1 e A.2 mostram os resultados das previsões para as duas classificações utilizando todas as variáveis do “Mapa da desigualdade”. A porcentagem se refere à frequência na coluna, assim é uma estimativa da probabilidade de acerto dado que foi previsto o risco daquela coluna. Esses modelos obtiveram respectivamente 80,03% ((691+156+75)/1152) e 87,5% ((811+197)/1152) de acerto de previsão.

Comparativamente, foi ajustado um modelo com apenas a variável índice de desigualdade de cada distrito. Os resultados estão nas Tabelas A.3 e A.4. O acerto de previsão nesses modelos foi de respectivamente 81,16% e 89,67%, portanto houve de fato uma aumento no acerto da previsão. Essa diferença com o teste de validação ocorreu porque anteriormente se estava considerando erro em uma previsão quantitativa e não de classes como se está fazendo agora. Dessa forma, combinado com o fato que esse modelo é mais simples que o outro por ter uma quantidade de variáveis muito menor, se escolheu utilizar apenas a variável índice de desigualdade, e não todo o conjunto de covariáveis que havia no estudo.

Até esse momento, nos modelos apresentados estavam incluídos também variáveis da quantidade casos da doença no mesmo distrito nos trimestres passados até 6 trimestres anteriores ao da variável resposta e a variável de quantidade de casos nos distritos vizinhos no trimestre anterior, já que a etapa de validação cruzada havia mostrado que elas diminuem o erro na previsão. Porém, agora que se optou por utilizar apenas o índice de desigualdade, foram testados subconjuntos dessas variáveis para verificar o quanto elas afetam o acerto na previsão.

Para todos os subconjuntos testados, o melhor era o que não continha a variável de número de casos no mesmo distrito dois trimestres atrás e a variável de número de casos nos distritos vizinhos no último trimestre. Assim, o modelo que apresentou os melhores resultados foi o modelo com as seguintes variáveis explicativas: número de casos da doença no mesmo distrito um trimestre atrás, número de casos da doença no mesmo distrito três trimestres atrás, número de casos da doença no mesmo distrito quatro trimestres atrás, número de casos da doença no mesmo distrito cinco trimestres atrás, número de casos da doença no mesmo distrito seis trimestres atrás e índice de desigualdade daquele distrito.

A análise demonstra que a inclusão de dados de trimestres anteriores a seis trimestres traz um ganho marginal na acurácia dos modelos, especialmente quando comparado ao uso de um intervalo mais recente (até três trimestres). A baixa contribuição entre eventos mais antigos e os casos atuais sugere uma contribuição limitada para a previsão, enquanto aumenta a complexidade do modelo.

Considerando o princípio de parcimônia - que prioriza modelos mais simples quando a diferença de desempenho é insignificante - optou-se por manter apenas as variáveis já validadas, garantindo um equilíbrio entre precisão e interpretabilidade.

As Tabelas A.5 e A.6 mostram os resultados de previsão desse modelo. O acerto de previsão dos modelos foi de 81,16% 90,36% respectivamente. A segunda classificação obteve previsões piores para todos os modelos, pois é mais difícil fazer uma previsão para 3 grupos diferentes. Aparentemente, o modelo com a primeira classificação se mostrou melhor, no entanto como foi dito, isso pode depender da escolha da pesquisadora de talvez fazer uma classificação que distingue mais a quantidade de casos por distrito, mesmo que isso prejudique a predição do modelo se isso for de interesse dela.

7. Conclusões

Este estudo investigou a dinâmica espaço-temporal da esporotricose felina em São Paulo entre 2011 e 2024, integrando georreferenciamento, variáveis socioeconômicas e

técnicas estatísticas para desenvolver um modelo preditivo de risco e técnicas de clusterização para identificação dos focos da doença. Os resultados revelam padrões críticos para o controle da doença:

1. **Crescimento exponencial e distribuição desigual**:

- A análise temporal confirmou um aumento acentuado de casos a partir de 2016, com picos após 2018.
- Espacialmente, observou-se concentração em distritos periféricos (e.g., Cangaíba, Sapopemba, São Mateus). A correlação negativa (Spearman: -0,665) entre casos e o *coeficiente de desigualdade* reforça que áreas socioeconomicamente vulneráveis são as mais afetadas.

2. **Focos de transmissão identificados**:

- A aplicação do DBSCAN detectou **28 focos ativos**, majoritariamente nas zonas leste e norte. Esses aglomerados irregulares (não limitados por divisões administrativas) evidenciam rotas de disseminação complexas.

3. **Desafios na modelagem preditiva**:

- Modelos temporais e GLM (*Generalized Linear Models*) testados apresentaram limitações na precisão, devido a:
  - **Heterogeneidade espacial**: A agregação de dados socioeconômicos por distrito (não por setor censitário) diluiu correlações locais críticas.

4. **Modelo preditivo**:

- O modelo preditivo não apresentou precisão tão alta, no entanto acredita-se que ele será capaz de entregar uma capacidade de tomada de decisões melhor para as autoridades da saúde, que era o interesse da pesquisadora.

APÊNDICE A

Tabelas

Tabela A.1 Previsões do modelo com todas as variáveis e primeira classificação

	Predito		
Real	**Risco Baixo**	**Risco Moderado**	**Risco Alto**
Risco Baixo	691 (91,28%)	88 (31,88%)	4 (3,36%)
Risco Moderado	63 (8,32%)	156 (56,21%)	40 (33,61%)
Risco Alto	3 (3,96%)	32 (11,59%)	75 (63,02%)

Tabela A.2 Previsões do modelo com todas as variáveis e segunda classificação

	Predito	
Real	**Risco Baixo**	**Risco Alto**
Risco Baixo	811 (92,68%)	80 (28,88%)
Risco Alto	64 (7,31%)	197 (71,11%)

Tabela A.3 Previsões do modelo com índice de desigualdade e primeira classificação

	Predito		
Real	**Risco Baixo**	**Risco Moderado**	**Risco Alto**
Risco Baixo	717 (89,29%)	65 (29,01%)	1 (0,8%)
Risco Moderado	85 (10,58%)	137 (61,16%)	37 (29,6%)
Risco Alto	1 (1,24%)	22 (9,82%)	87 (69,6%)

Tabela A.4 Previsões do modelo com índice de desigualdade e segunda classificação

	Predito	
Real	**Risco Baixo**	**Risco Alto**
Risco Baixo	839 (92,6%)	52 (21,13%)
Risco Alto	67 (7,39%)	194 (78,86%)

Tabela A.5 Previsões do modelo final primeira classificação

	Predito		
Real	**Risco Baixo**	**Risco Moderado**	**Risco Alto**
Risco Baixo	717 (88,73%)	65 (29,54%)	1 (0,8%)
Risco Moderado	90 (11,13%)	132 (60%)	37 (29,83%)
Risco Alto	1 (1,23%)	23 (10,45%)	86 (69,35%)

Tabela A.6 Previsões do modelo final segunda classificação

	Predito	
Real	**Risco Baixo**	**Risco Alto**
Risco Baixo	844 (92,95%)	47 (19,26%)
Risco Alto	64 (7,04%)	197 (80,73%)

APÊNDICE B

Figuras

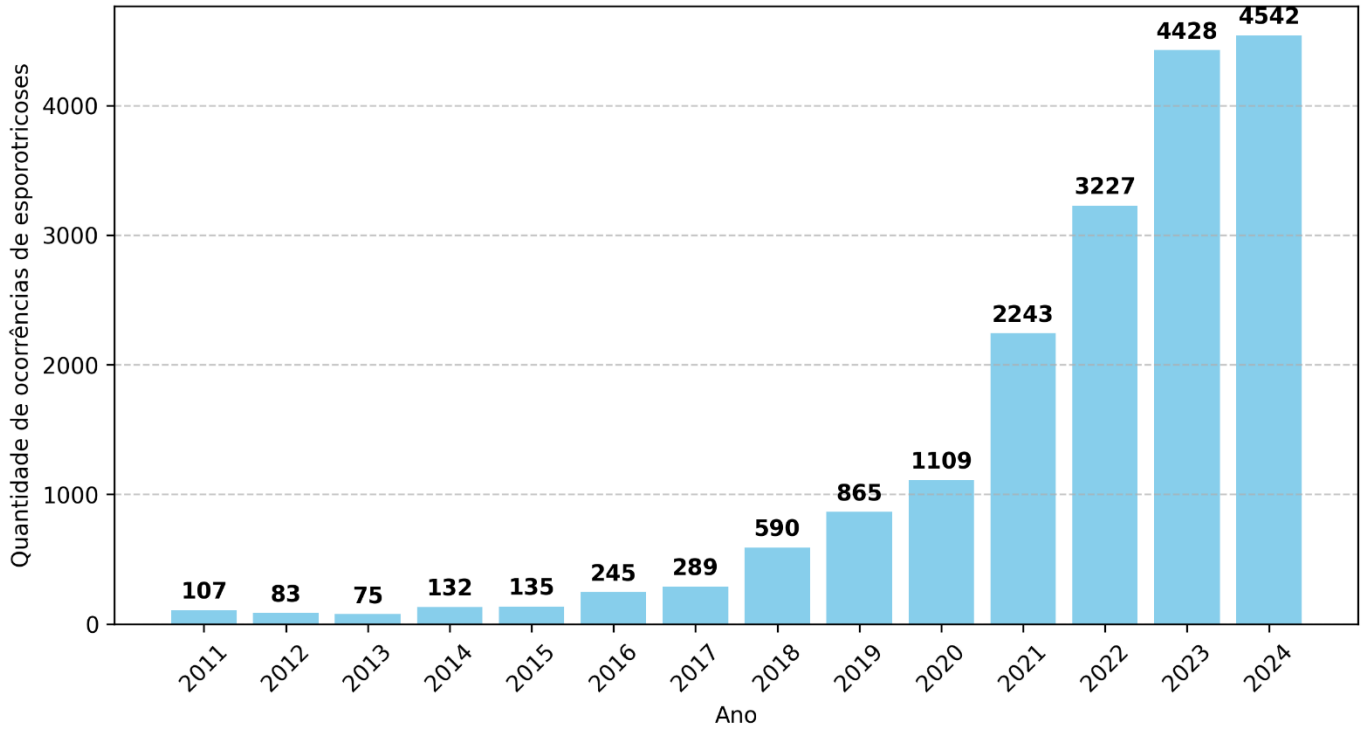

Figura B.1 Quantidade de ocorrências de esporotricose agregada por ano.

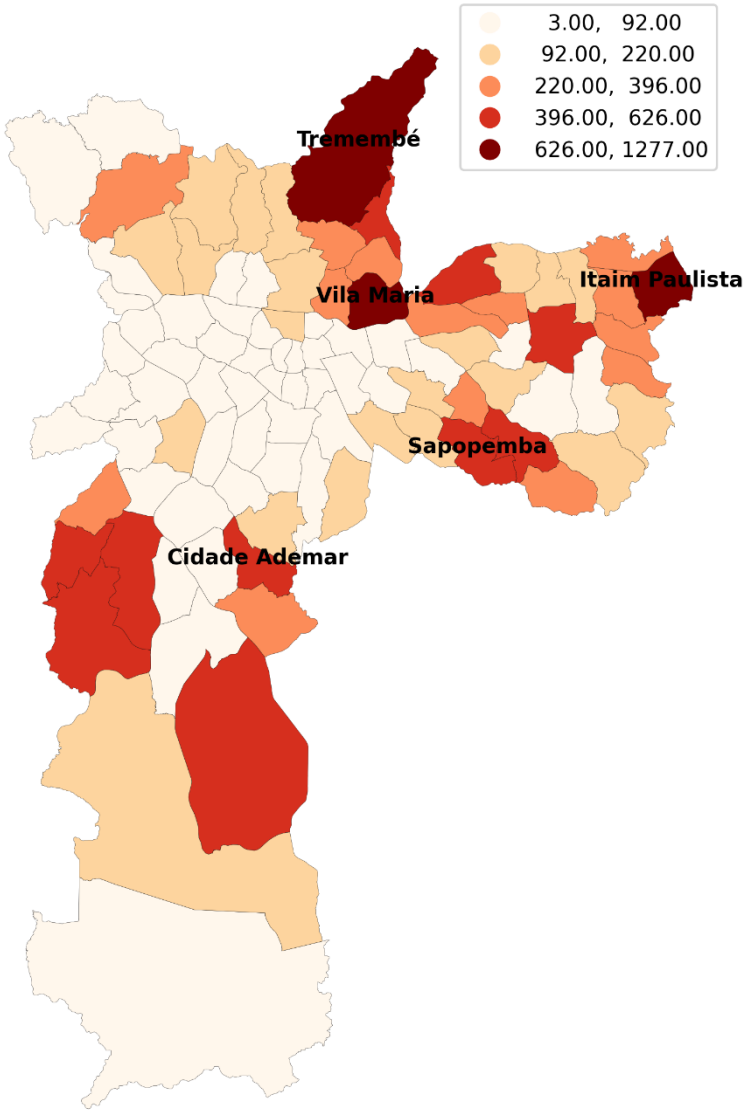

Figura B.2 Mapa da quantidade de ocorrências de esporotricose agregada por distrito.

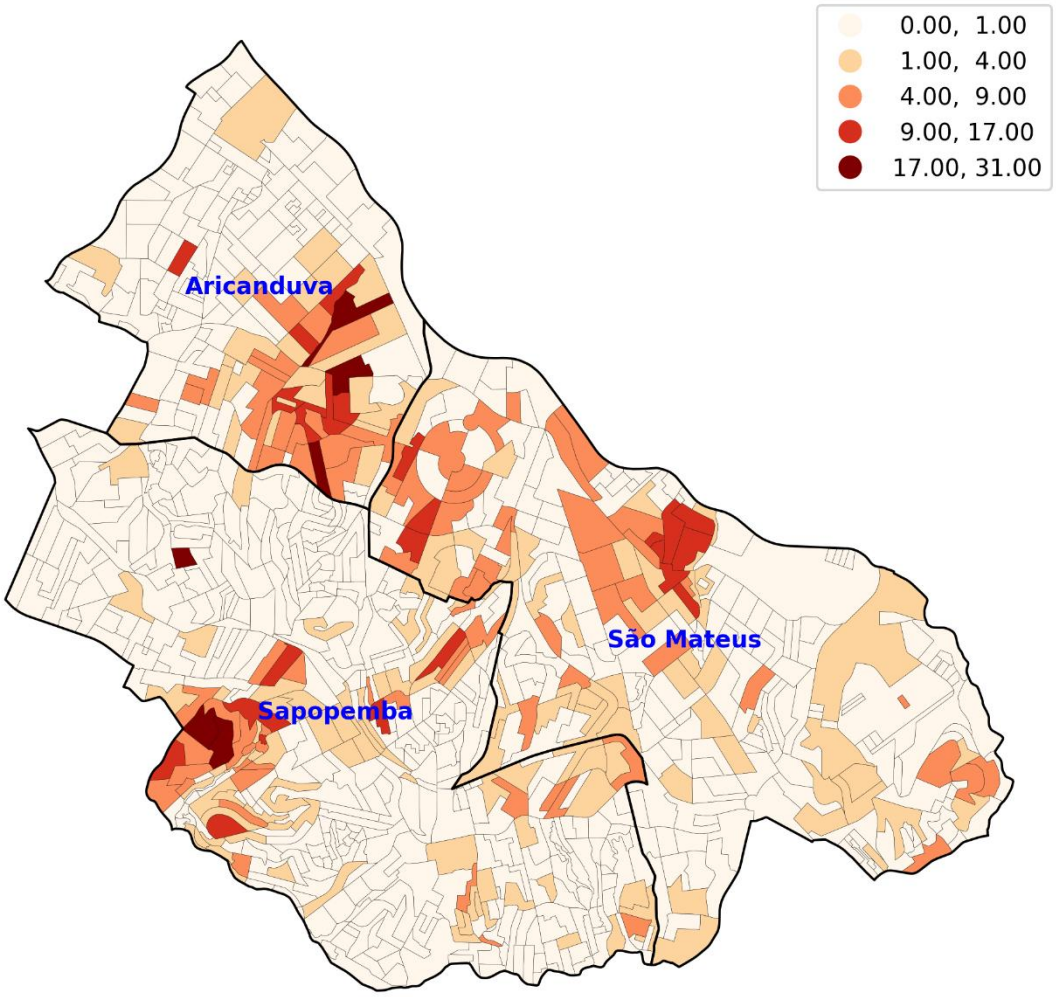

Figura B.3 Mapa da quantidade de ocorrências de esporotricose agregada por setor censitário para alguns distritos.

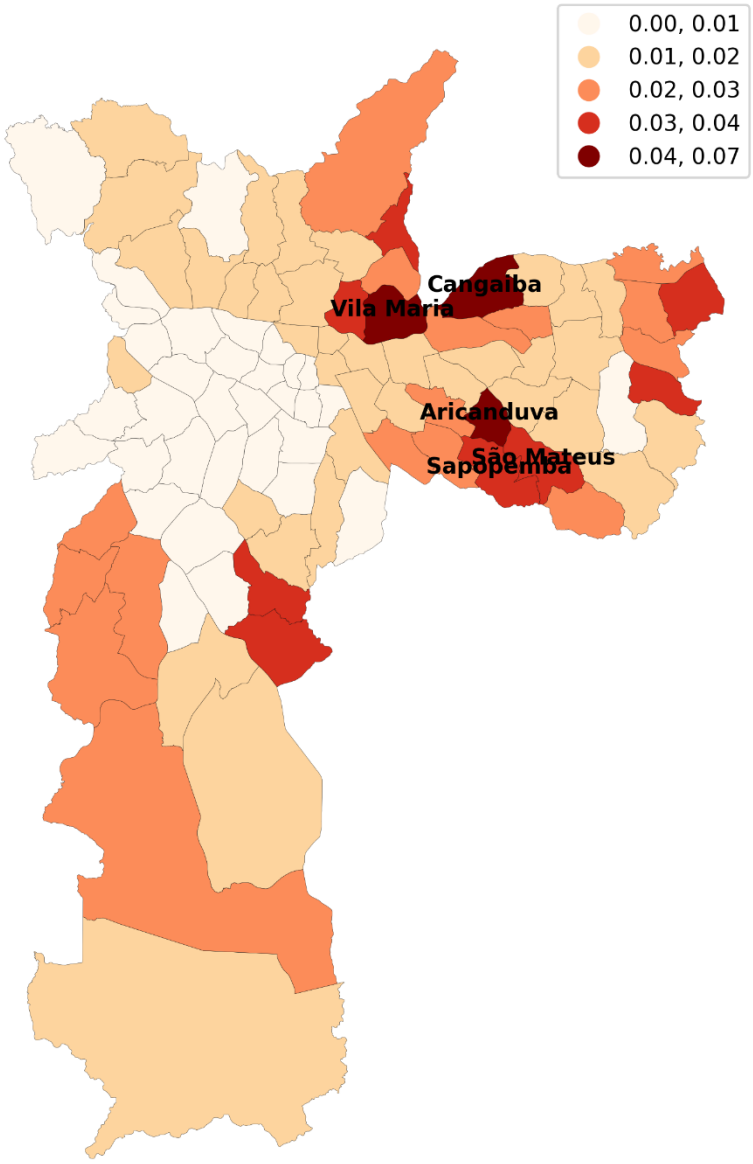

Figura B.4 Mapa da proporção de ocorrências de esporotricose pela quantidade de gatos por distrito.

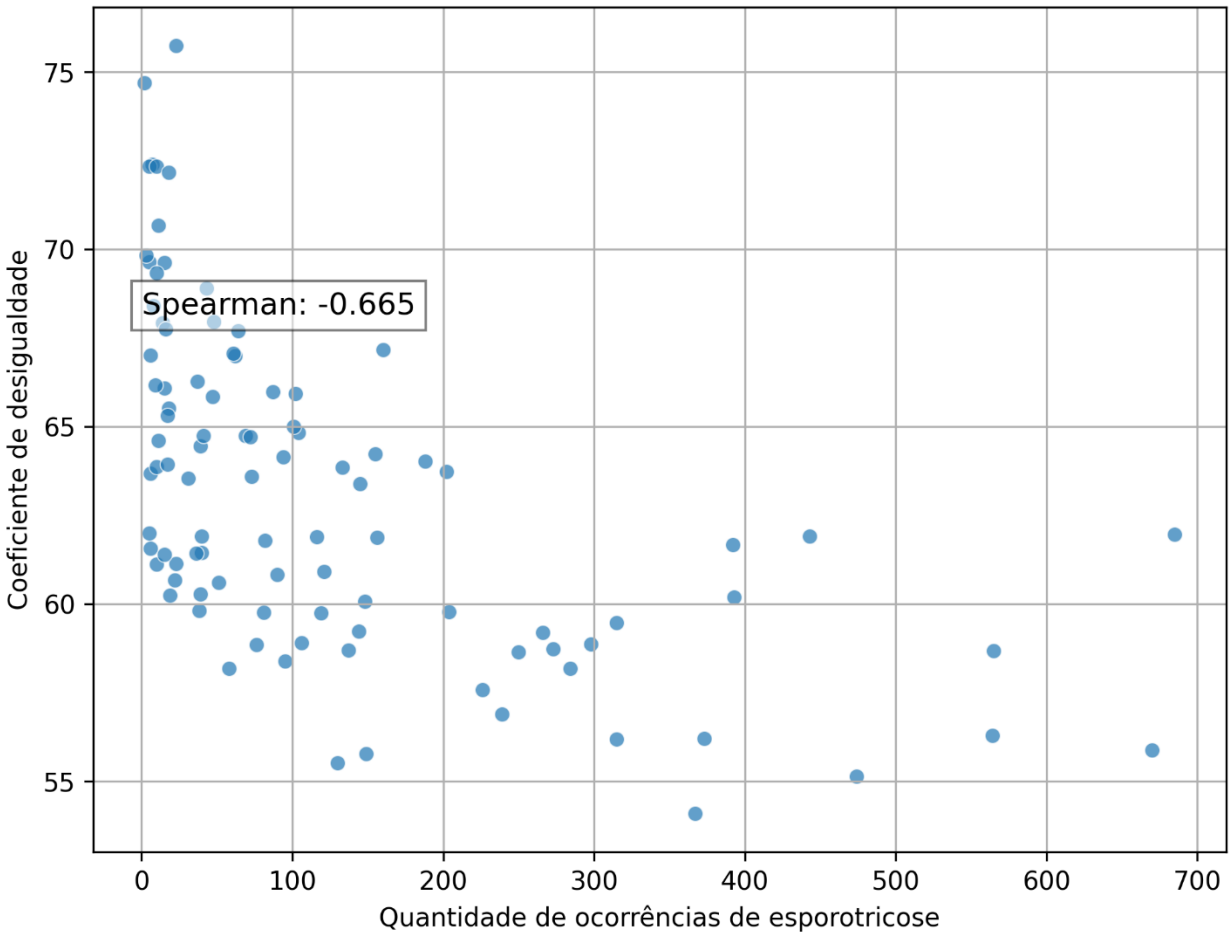

Figura B.5 Diagrama de dispersão entre a quantidade de ocorrências de esporotricose e o coeficiente de desigualdade, por distrito.

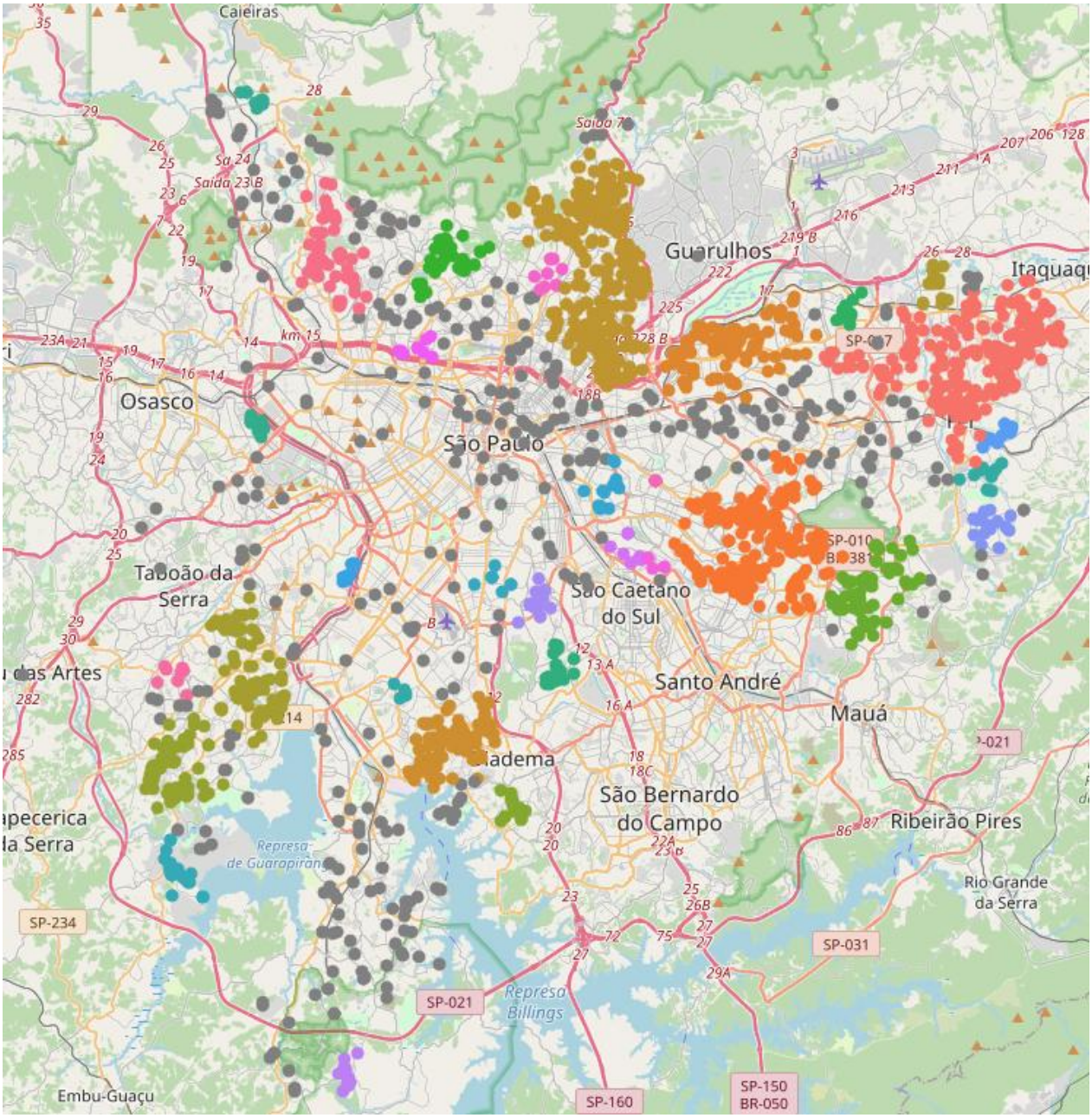

Figura B.6 Identificação de focos da doença segundo o método DBSCAN