UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Botucatu

Vinícius Narciso Fernandes

Análise descritiva sobre dados de dengue com relação aos aspectos ambientais, sociodemográficos e geográficos do Estado de São Paulo

Botucatu, São Paulo

2023

Vinícius Narciso Fernandes

Análise descritiva sobre dados de dengue com relação aos aspectos ambientais, sociodemográficos e geográficos do Estado de São Paulo

Trabalho de conclusão de curso apresentado ao curso de Física Médica da Universidade Estadual Paulista "Júlio de Mesquita Filho" - Instituto de Biociências de Botucatu como parte dos requisitos necessários para obtenção do título de Bacharel em Física Médica.

Orientadora: Profa. Dra. Cláudia Pio Ferreira
Co-orientadores: Thomas Vilches e Wesley Cota

Botucatu, São Paulo
2023

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: MARIA CAROLINA A. CRUZ E SANTOS-CRB 8/10188

Fernandes, Vinícius Narciso.
Análise descritiva sobre dados de dengue com relação aos aspectos ambientais, sociodemográficos e geográficos do Estado de São Paulo / Vinícius Narciso Fernandes. - Botucatu, 2023

Trabalho de conclusão de curso (bacharelado - Física Médica) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Cláudia Pio Ferreira
Coorientador: Thomas Nogueira Vilches
Coorientador: Wesley Cota
Capes: 10104003

1. Análise de componentes principais. 2. Dengue. 3. Correlação (Estatística).

Palavras-chave: Análise de componentes principais; Análise descritiva; Medidas de correlação.

Vinícius Narciso Fernandes

Análise descritiva sobre dados de dengue com relação aos aspectos ambientais, sociodemográficos e geográficos do Estado de São Paulo

Trabalho de Conclusão de Curso defendido e aprovado em Botucatu, 12 de dezembro de 2023, pela banca examinadora constituída pelos professores: Luzia Aparecida Trinca e Fernando Luiz Pio dos Santos.

Dedicatória

Agradeço a professora Cláudia Pio pela oportunidade de trabalho, evolução e maturidade científica que pude obter ao longo desse semestre. Agradeço ao Thomas Vilches e ao Wesley Cota por toda ajuda e orientação que também me deram ao longo do semestre no presente trabalho.

Agradecimentos

Agradeço inicialmente à Professora Cláudia Pio, por me proporcionar a oportunidade de realizar este trabalho acreditando em mim num momento onde estava me levantando como pessoa. Quero agradecer também ao Thomas Vilches e Wesley Cota que nos últimos tempos sempre vem com relação ao trabalho.

Aos amigos que fiz durante esses anos na faculdade, em especial, Garibaldo, Emotiva, América, Tijolin, Hello Kitty, Parabéns, Pet, Ana Flávia e Suzana Wesselka assim como vários outros que foram aparecendo no decorrer dos anos, que enfim são tantos que fica difícil de agradecer a todos, obrigado a vocês que me ajudaram nesses últimos anos nessa etapa da minha vida e me proporcionaram momentos de alegria. E também, agradeço à toda *Física Médica XV*.

Aos meus avós, tanto maternos quanto paternos, dedico a minha formação a eles que me ajudaram em momentos difíceis, especialmente ao meu avô Antônio espero que de onde ele esteja possa estar feliz por mim e orgulhoso de conquistar a faculdade que queria e de conseguir me levantar a pesar de todos os percalços do meio do caminho. Ter me tornado o homem diferente do que ele achava que eu poderia ser.

Quero também agradecer a todos os professores que tive durante a minha jornada até a faculdade, foram esses que sempre me incentivaram a estudar e a ir buscar o caminho do conhecimento me mostrando que através da educação poderia me tornar uma pessoa melhor e com perspectivas de futuro. Realmente a única coisa que posso lhes dizer é muito obrigado por me guiarem no caminho correto.

Agradecer a própria Unesp e todo o seu corpo estudantil que sem eles seria difícil fazer tudo sozinho. A estrutura de estudo que me ofereceu e ao apoio financeiro por meio do auxílio socioeconômico que através desse benefício também pude me manter na faculdade, assim me deixando mais tranquilo para focar nos estudos e menos preocupado com questões econômicas.

Este trabalho foi financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), com suporte computacional do Laboratório de Epidemiologia de Doenças Infecciosas do Departamento de Infectologia (LEDI) da Faculdade de Medicina de Botucatu/Unesp, e dados fornecidos pelo Centro Conjunto Brasil-Reino Unido para Descoberta, Diagnóstico, Genômica e Epidemiologia de Arbovírus (CADDE).

”Meu cérebro é apenas um receptor, no Universo existe um núcleo a partir do qual obtemos conhecimento, força e inspiração. Eu não penetrei nos segredos deste núcleo, mas eu sei que ele existe”

– Nikola Tesla

Resumo

A dengue é uma arbovirose de importância mundial. Não existe vacina e o controle é feito sobre a população de vetores, o mosquito *Aedes*. A definição de cidades alvo para o investimento em saúde pública é desejável. Este trabalho tem como objetivo estudar como as cidades do estado de São Paulo se agrupam de acordo com suas características sociodemográficas e climáticas, e a relação destas com a incidência de dengue nestas cidades. Para isso, dados de temperatura, umidade e pluviosidade foram utilizados para a obtenção do índice P, que mede o potencial de transmissão de arboviroses. A essas variáveis juntou-se também outras sociodemográficas. Através de uma análise de componentes principais, foi feita uma redução da dimensionalidade dos dados, e, após, aplicada técnicas de agrupamento hierárquica e não-hierárquica. Os grupos formados apresentaram padrões de incidência de dengue distintos, sendo possível selecionar dentre eles, aqueles com maior número de casos. Os métodos aqui apresentados podem auxiliar na identificação de cidades alvo para a implementação de políticas públicas para controle da transmissão de arboviroses.

Palavras-chave: Análise Descritiva, Medida de Correlação, Análise de componentes Principais.

Abstract: Dengue disease is an arbovirusis of global importance. There is no vaccine and control is carried out over the vector population, the *Aedes* mosquito. The definition of target cities for investment in public health is desirable. This work aims to study how cities in the state of São Paulo are grouped according to their sociodemographic and climatic characteristics, and their relationship with the incidence of dengue in these cities. For this, temperature, humidity and rainfall data were used to obtain the P index, which measures the potential for transmission of arboviruses. These variables were also joined by other sociodemographic variables. Through a principal component analysis, the dimensionality of the data was reduced, and then hierarchical and non-hierarchical clustering techniques were applied. The groups created showed different dengue incidence patterns, making it possible to select among them those with the highest number of cases. The methods presented here can help to identify target cities for implementing public policies to control the transmission of arboviruses.

Key-words: Descriptive Analysis, Correlation, Principal Component Analysis.

Conteúdo

1 Introdução

A dengue é uma doença febril aguda, sistêmica e dinâmica - o paciente pode evoluir de um estágio para outro rapidamente - e apresenta um amplo espectro clínico, de casos assintomáticos a graves que podem evoluir para óbito. É uma arbovirose cujo principal vetor é o mosquito *Aedes aegypti* e o agente etiológico é um vírus da família *Flaviviridae*. Quatro diferentes sorotipos do vírus (DENV-1, DENV-2, DENV-3 e DENV-4) causam a infecção e indivíduos recuperados tem imunidade permanente ao vírus homólogo, e temporária ao vírus heterólogo de maneira que reinfecções são observadas [1]. Todas as faixas etárias são igualmente suscetíveis à doença, contudo, idosos e pessoas com doenças crônicas, como diabetes e hipertensão arterial, têm maior risco de evoluir para casos graves e complicações [2].

A dengue clássica é autolimitada, dura em torno de 7 a 10 dias, e os sintomas mais comuns são febre alta (39 e 40 graus Celsius), mialgia e cefaleia. A dengue hemorrágica acontece, geralmente, durante a infecção secundária, mas pode ocorrer também em infecções primárias, especialmente em lactentes. Em casos leves e moderados, a febre diminui com sudorese profunda. Também podem ser observadas pequenas alterações na frequência do pulso e pressão arterial, com extremidades frias e edema. Já em casos graves, pode haver agravamento súbito após alguns dias com progresso para Síndrome do Choque da Dengue (SCD) [3].

No Brasil a doença é considerada um problema de saúde pública, e as condições socioambientais atreladas à baixa efetividade de programas de combate ao vetor causam cada vez mais preocupação [4]. O período do ano de maior transmissão da doença ocorre nos meses mais chuvosos (ou meses úmidos) de cada região do país, e geralmente se inicia em novembro e vai até maio [2]. A expansão das áreas de ocorrência de dengue no Brasil está associada à urbanização e à concentração demográfica sem uma devida estrutura de saneamento básico. Outros fatores como as alterações climáticas, impactos ambientais nas paisagens e ecossistemas, predomínio de novos modelos e estilos de vida da população também impactam a dinâmica da transmissão da doença [5]. Em resumo, fatores antropogênicos e climáticos contribuem não somente para a dispersão ativa do vetor como também para a disseminação de vários sorotipos da doença [6].

Esta monografia descreve o trabalho de iniciação científica realizado de junho de 2023

a dezembro de 2023 pelo aluno Vinícius Narciso Fernandes no Departamento de Biodiversidade e Bioestatística , com a Profa. Dra. Cláudia Pio Ferreira e co-orientado por Thomas Nogueira Vilches e Wesley Cota, no tema de análise de dados. Durante o estágio o aluno se familiarizou com o processador de texto LaTeX e aprimorou seus conhecimentos na Linguagem de Programação R. É bolsista PIBIC-CNPq desde Outubro de 2023 sob a responsabilidade de Profa. Dra. Claúdia Pio Ferreira. Resultados parciais do trabalho foram apresentados em congressos, (Congrebio) - XII Congresso de Biociências - em Agosto de 2023 (Botucatu-SP), CIC - Congresso de Iniciação Científica - em Outubro de 2023 (Botucatu - SP).

2 Objetivo

Estudar a associação entre variáveis climáticas, demográficas e socioeconômicas com a incidência de dengue nas diferentes cidades do Estado de São Paulo.

Objetivos específicos:

- Calcular o índice P, potencial de transmissão do vírus, o qual depende dos fatores abióticos (temperatura e umidade).
- Executar uma análise de componentes principais.
- Executar uma análise de agrupamento entre as cidades conforme as características climáticas, demográficas e socioeconômicas.
- Analisar a incidência de dengue entre os diferentes grupos e sua associação com as variáveis de interesse.

3 Metodologia

3.1 Índice P

O potencial de transmissão de um patógeno pode ser medido pelo número de reprodução básico (R_0) ou efetivo (R_e). O R_0 do patógeno mede o número de casos secundários gerado, em média, por um único hospedeiro infectado que chega em uma

população totalmente suscetível. No caso de vírus transmitido por mosquito, R_0 é dado pela soma do potencial reprodutivo (transmissão) de cada mosquito fêmea adulta, $P_{(u,t)}$, onde u será a umidade e t a temperatura, multiplicado pela razão entre o número total de mosquitos fêmeas e o número total de indivíduos humanos, M. Já o R_e é medido/interpretado de maneira similar, mas leva em consideração a presença de hospedeiros imunes (S_h e S_v medem, respectivamente, o número de indivíduos humanos e vetores suscetíveis em cada instante de tempo), o que dificulta transmissão da doença [5],

$$R_0 = MP_{(u,t)}, \quad R_e = R_0 S_h S_v \,, \tag{1}$$

com

$$P_{(u,t)} = \frac{a_u^\nu \, \phi_{(t)}^{\nu\to h} \, \phi^{h\to\nu} \, \gamma_{(t)}^\nu \gamma^h}{\mu_{(u,t)}^\nu (\sigma^h + \mu^h)\,(\gamma^h + \mu^h)\,(\gamma_{(t)}^\nu + \mu_{(u,t)}^\nu)} \,. \tag{2}$$

Neste modelo, há um total de oito parâmetros na expressão de R_0, quatro não dependem de fatores abióticos (o tempo de vida humano $1/\mu^h$, a probabilidade de transmissão do humano infectado para o mosquito $\phi^{h\to\nu}$, o período infeccioso humano $1/\sigma^h$ e tempo de incubação intrínseco do vírus $1/\gamma^h$) e quatro dependem de fatores abióticos (o tempo de vida dos mosquito adultos $1/\mu_{(u,t)}^\nu$, o tempo de incubação extrínseco $1/\gamma_{(t)}^\nu$, a taxa de picada diária $a_{(u)}^\nu$ e a probabilidade de transmissão do mosquito infectado para o indivíduo humano $\phi_{(t)}^{\nu\to h}$). Os parâmetros influenciados pelos fatores abióticos (umidade e temperatura), são funções previamente determinas em estudos experimentais por estimativas laboratoriais de dados entomológicos do mosquito sob diversas condições de temperatura e umidade. Portanto, $R_0 := R_0(u,t)$, $R_e := R_e(u,t)$, $S_h := S_h(t)$ e $S_v := S_v(t)$.

3.2 Correlação linear

A correlação linear é uma medida estatística que informa se existe uma relação linear entre duas variáveis, X_1 e X_2. Para isso é calculado o coeficiente de correlação (ρ_{12}) que indica a força e a direção da relação entre essas variáveis. Dadas J variáveis $X_j (j = 1,2,3,\cdots,k)$ de interesse, cada uma com n observações, podemos definir a matriz de correlações ρ como [7]

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \dots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \dots & 1 \end{bmatrix}, \tag{3}$$

a qual é simétrica em relação à diagonal principal que, tem valores iguais a 1. Dadas as variáveis X_i e X_j, o coeficiente de correlação linear ρ_{ij} pode ser calculado com base na expressão

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \times \sigma_j} \tag{4}$$

em que $\overline{X}_i$ e $\overline{X}_j$ representam, respectivamente, os valores médios das variáveis X_i e X_j, dados por

$$\overline{X}_j = \frac{\sum_{m=1}^{n} X_{jm}}{n},\ j = 1,...,k. \tag{5}$$

Com relação à força do coeficiente de correlação linear tem-se que [8]:

- Se $|\rho| \geq 0.9$ a correlação é muito forte;
- Se $|\rho| \in (0.7; 0.9)$ a correlação é forte;
- Se $|\rho| \in (0.5; 0.7]$ a correlação é moderada;
- Se $|\rho| \in (0.3; 0.5]$ a correlação é fraca; e
- Se $|\rho| < 0.3$ a correlação é desprezível.

A correlação pode ser direta ou inversamente proporcional. Dessa forma, se o valor da correlação é positivo, a relação é direta, ou seja, quando uma variável aumenta a outra também aumenta. Se a correlação for negativa, isso indica que as variáveis são inversamente proporcionais, ou seja, enquanto uma variável aumenta a outra diminui.

3.3 Análise de Componentes Principais

A análise de componentes principais (ACP) é uma técnica estatística de análise multivariada que transforma linearmente um conjunto original de variáveis, inicialmente correlacionadas entre si, num conjunto de variáveis de mesma dimensão denominada de

componentes principais. Esse novo conjunto tem propriedades importantes como: cada componente principal é uma combinação linear de todas as variáveis originais; as novas variáveis formadas são independentes entre si e possuem o máximo de informação em termos da variação total contida nos dados [7]. Essa técnica pode ser utilizada para reduzir o número de dimensões dos dados originais com a menor perda possível de informação [9].

Entre as vantagens tem-se:

- Retirada da multicolinearidade das variáveis, uma vez que permite transformar um conjunto de variáveis originais inter-relacionadas em um novo conjunto de variáveis não correlacionadas (Componentes Principais).
- Reduzir muitas variáveis a eixos que representam algumas variáveis (Componentes principais), sendo estes eixos perpendiculares (ortogonais) explicando a variação dos dados de forma decrescente.

Entre as desvantagens tem-se:

- Se baseia em relações lineares, então, se as relações obedecem outra métrica não será útil;
- São sensíveis a dados com valores discrepantes;
- Não é adequada o seu uso dados ausências;
- Não é útil o uso da técnica quando o número de variáveis é maior do que o número de observações.
- Não é possível o uso destas na construção de modelos de predição.

Sejam as variáveis X_1,X_2,$\cdots$,X_J, cada um com n medidas. Este conjunto pode ser representado em um matriz $X(n \times k)$ [10]:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}, \tag{6}$$

e tem matriz de covariância Σ dada por

$$\Sigma = \begin{bmatrix} \sigma_{11}, & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{bmatrix}, \quad \text{com} \quad \sigma_{ij} = \sum_{m=1}^{n} \frac{(X_{im} - \overline{X}_i)(X_{jm} - \overline{X}_j)}{(n-1)}. \tag{7}$$

Calcula-se então os autovalores e autovetores da matriz Σ, que é semi-positiva definida, o que garante a não-negatividade de seus autovalores. Encontram-se os pares de autovalores e autovetores $(\lambda_1,e_1),(\lambda_2,e_2),\cdots,(\lambda_k,e_k)$, em que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$, associados à matriz de covariância. Os autovetores tem k elementos, i.e., $e_i = (e_{i1},e_{i2},...,e_{ik})$, portanto o i-ésimo componente principal é definido por [10]

$$Z_i = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ik}X_k\,. \tag{8}$$

Utilizando a decomposição espectral da matriz de covariância (matriz simétrica), dada por $\Sigma = P\Lambda P^t$, em que P é a matriz ortonormal, i.e., $P^t = P^{-1}$, composta pelos autovetores de Σ em suas colunas, e Λ é a matriz diagonal formada pelos autovalores de Σ, tem-se que

$$tr(\Sigma) = tr(P\Lambda P^t) = tr(\Lambda P^{-1}P) = tr(\Lambda I) = tr(\Lambda) = \sum_{i=1}^{k} \lambda_i\,, \tag{9}$$

portanto, a variabilidade total contida nas variáveis originais é igual a soma dos autovalores contida nos componentes.

A contribuição de cada componente principal (Z_i) é expressa em porcentagem, e a explicação individual de cada componente pode ser calculada, da seguinte forma:

$$C_i = \frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i} \times 100 = \frac{\lambda_i}{tr(\Sigma)} \times 100\,. \tag{10}$$

Para a seleção das componentes principais a serem utilizadas, pode-se aplicar o critério de Kaiser (ou critério das raízes latentes). Com esse critério retêm-se as componentes principais com autovalores maiores do que a unidade ($\lambda_i > 1$) e garante-se que essas explicam a maior parte da variação dos dados.

3.4 Análise de agrupamentos

A análise de agrupamentos representa um conjunto de técnicas exploratórias que podem ser aplicadas quando há a intenção de se verificar a existência de comportamentos

semelhantes entre observações em relação a determinadas variáveis, visando a criação de grupos de forma que os elementos de um mesmo grupo sejam homogêneos e os elementos em grupos diferentes sejam heterogêneos. Para isso, uma medida de distância ou de semelhança, que servirá de base para que as observações sejam consideradas menos ou mais próximas, é escolhida, assim como um algoritmo de formação de grupos, que deverá ser definido entre os métodos hierárquicos e não hierárquicos [7].

Os métodos hierárquicos permitem a identificação do ordenamento e da alocação das observações, oferecendo possibilidades para que o pesquisador estude, avalie e decida sobre o número de grupos a serem formados. Já nos métodos não-hierárquicos, parte-se de uma quantidade conhecida de grupos e, a partir de então, é elaborada a alocação das observações nesses grupos, com posterior avaliação da representatividade de cada variável na formação dos grupos. A formação de grupos é bastante sensível à presença de valores discrepantes, e a exclusão ou a retenção desses valores na base de dados depende dos objetivos de pesquisa e da natureza dos dados [7].

As técnicas de análise de agrupamentos são consideradas exploratórias, ou de interdependência, uma vez que suas aplicações não apresentam caráter preditivo para outras observações não presentes inicialmente na amostra. Assim, os métodos de análise de agrupamentos são chamados de procedimentos estáticos e não-supervisionados, uma vez que a inclusão de novas observações no banco de dados torna necessária a reaplicação da modelagem para que, sejam gerados novos agrupamentos [7].

Inicialmente é feito uma padronização das variáveis. O método padrão é o procedimento de *z-scores*, em que, para cada observação i, o valor de uma nova variável padronizada ZX_j é obtido pela subtração do correspondente valor da variável original X_j, pela sua média $\overline{X}_j$ e, prosseguindo, o valor resultante é dividido pelo desvio-padrão s_j.

$$Z_{jm} = \frac{X_{jm} - \overline{X}_j}{s_j} \quad \text{com} \quad s_j = \sqrt{\frac{\sum_{i=1}^{n}(X_{ji} - \overline{X}_j)^2}{N-1}}. \tag{11}$$

Tal procedimento é necessário, uma vez que as variáveis podem apresentar magnitudes dos valores e da natureza das unidades de medida muito distintas; todas as respectivas variáveis padronizadas pelo procedimento citado acima terão média igual a 0 e desvio-padrão igual a 1.

3.4.1 Agrupamentos hierárquicos e não-hierárquicos

Agrupamento hierárquico é uma técnica que pode ser classificada em aglomerativa ou divisiva. No método aglomerativo, inicialmente, cada elemento é considerado ser um grupo individual e ao longo das etapas os elementos vão se agrupando até que no fim exista somente um grupo com todos os elementos. O método divisivo consiste em considerar todos os elementos inicialmente em um único grupo e, ao longo das etapas, os grupos vão se dividindo, até que na última etapa cada grupo terá um único elemento. Em geral, os métodos aglomerativos exigem uma capacidade computacional menor que os divisivos [7].

A medida de distância (dissimilaridade) mais comum é a distância euclidiana, embora, dependendo do tipo de dados, existam outras métricas como a distância quadrática euclidiana, Minkowski, Manhattan, Chebychev e Canberra, isso dependendo dos dados. A distância euclidiana pode ser definida como:

$$d_{pq} = \sqrt{\sum_{j=1}^{k} (ZX_{pj} - ZX_{qj})^2} \tag{12}$$

onde ZX_{jp} representa a característica do indivíduo p, ZX_{jq} representa a característica do indivíduo q, e k representa o número de variáveis na amostra.

Entre os esquemas hierárquicos aglomerativos, cita-se o método de encadeamento. Nesse, dois grupos sofrem fusão com base na distância média entre todos os pares de observações pertencentes a esses grupos, a qual é dada por

$$d_{(MN)W} = \frac{\sum_{p=1}^{m+n} \sum_{q=1}^{w} d_{pq}}{(m+n)w} \tag{13}$$

em que d_{pq} representa a distância entre qualquer observação p do agrupamento MN e qualquer observação q do agrupamento W, e $m+n$ e w representam, respectivamente, a quantidade de observações nos agrupamentos MN e W.

Dentre os esquemas de aglomeração não-hierárquicos, a técnica do *K-means* é amplamente utilizada. Nesta, a quantidade de grupos K é definida de maneira prévia pelo pesquisador. A técnica busca minimizar as distâncias entre os elementos e o centroide do grupo a que ele é atribuído [11].

3.5 Descrição dos Dados

Os presente trabalho utilizou quatro banco de dados: (i) incidência de dengue, (ii) características demográficas da população, (iii) características geográficas das cidades, e (iv) dados climáticos. O banco formado é referente a todas as cidades do estado de São Paulo e contempla o período de 2007 a 2019. Neste trabalho consideramos apenas o período de 01 de janeiro de 2015 até 31 de dezembro de 2019.

O banco de dados de incidência de dengue contém informações anonimizada de mais de quatro milhões de notificações de casos de dengue. Esse banco é um recorte da base de dados de Doenças de Agravo e Notificação do Ministério da Saúde feita pela Secretaria de Saúde do Estado de São Paulo. Informações como data de notificação, data de primeiros sintomas, evolução do caso, idade, sexo, cor, nível de educação formal, cidade de notificação e residência constam nesse banco. As bases de dados sobre características demográficas da população e da cidade foram obtidas do censo demográfico do IBGE de 2010, e os dados climáticos das cidades do Estado de São Paulo foram obtidos do Instituto Nacional de Meteorologia. Os bancos de dados foram acessados por meio do sistema computacional do Laboratório de Epidemiologia de Doenças Infecciosas do Departamento de Infectologia (LEDI) da Faculdade de Medicina de Botucatu, coordenado pelo Prof. Dr. Carlos M. C. B. Fortaleza (FMB-Unesp). A série temporal de dengue foi obtida por intermédio da Profa. Dra. Ester Sabino (IMT-USP), e tratados pelo Dr. Wesley Cota (IMT-USP & FMB-Unesp) e fazem parte de um banco de dados maior pertencente ao Centro Conjunto Brasil-Reino Unido para Descoberta, Diagnóstico, Genômica e Epidemiologia de Arbovírus (CADDE).

4 Resultados e Discussão

4.1 O índice P

A Figura 1 mostra a média do índice P em cada município do estado de São Paulo no período de 01 de janeiro de 2015 a 31 de dezembro de 2019 e a incidência acumulada de dengue observada no mesmo período.

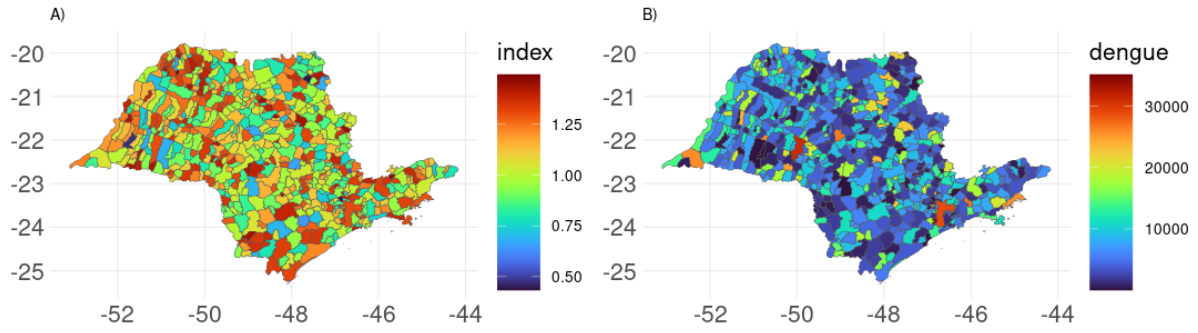

Figura 1: (A) Média do índice potencial de transmissão de arboviroses no Estado de São Paulo, e (B) Incidência de dengue acumulada por 100 mil habitantes. Os dados contemplam o período de 01 de janeiro de 2015 a 31 de dezembro de 2019.

Cores quentes (amarelo e vermelho) indicam os municípios com potencial de transmissão de dengue maior, enquanto as frias (azul e verde) mostram os locais com índice P menor e, portanto, menor potencial de transmissão da dengue. Podemos observar uma maior concentração de cidades com índice P maior do que 1 na região oeste e noroeste do estado de São Paulo.

Figura 2 mostra a relação não linear entre a incidência acumulada de dengue e o índice P médio. Observa-se que conforme o potencial de transmissão de arboviroses aumenta, a incidência acumulada de dengue aumenta também (o mesmo padrão é observado se utilizarmos a média da incidência de dengue no período).

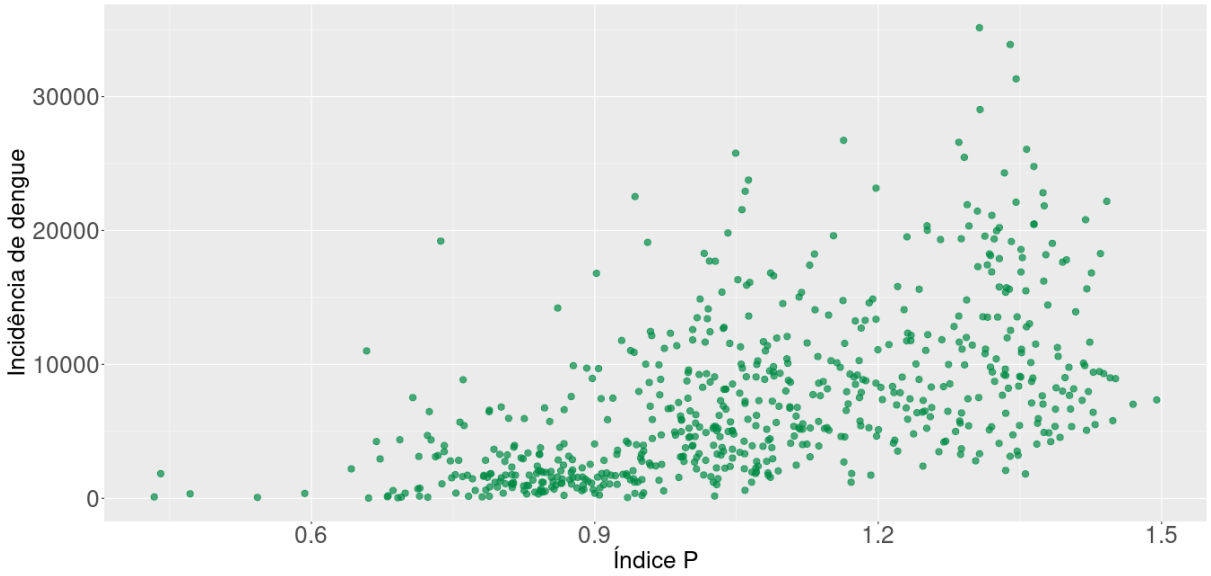

Figura 2: Incidência acumulada de dengue por 100 mil habitantes em função do índice P médio, calculado no período de 01 de janeiro de 2015 a 31 de dezembro de 2019.

4.2 Correlação entre as variáveis

As variáveis climáticas estudadas foram: temperatura (T), umidade (H) e pluviosidade (R). Utilizou-se para cada uma dessas variáveis, as médias dos valores máximos, mínimos e médios mensais observados em cada cidade do Estado de São Paulo durante o período estudado (01/2015 a 12/2019). Já as variáveis demográficas selecionadas foram: número médio de indivíduos por domicílios, renda média por domicílios, quantidade média de esgoto gerado por domicílios, quantidade média de lixo gerado por domicílios e quantidade de água média consumida por domicílios. Acrescentou-se o potencial de transmissão de arboviroses (P) diário a esse estudo.

Como mostrado na Figura 3, a temperatura (média, máxima e mínima) apresenta uma correlação forte e positiva com o índice P (máximo e mínimo). Já outras variáveis como, pluviosidade e umidade, a correlação entre elas e o índice P é forte e negativa. Observa-se ainda que as variáveis demográficas apresentam correlação fraca com as variáveis climáticas, como esperado. Contudo, as variáveis sociodemográficas apresentam correlação positiva muito forte entre si.

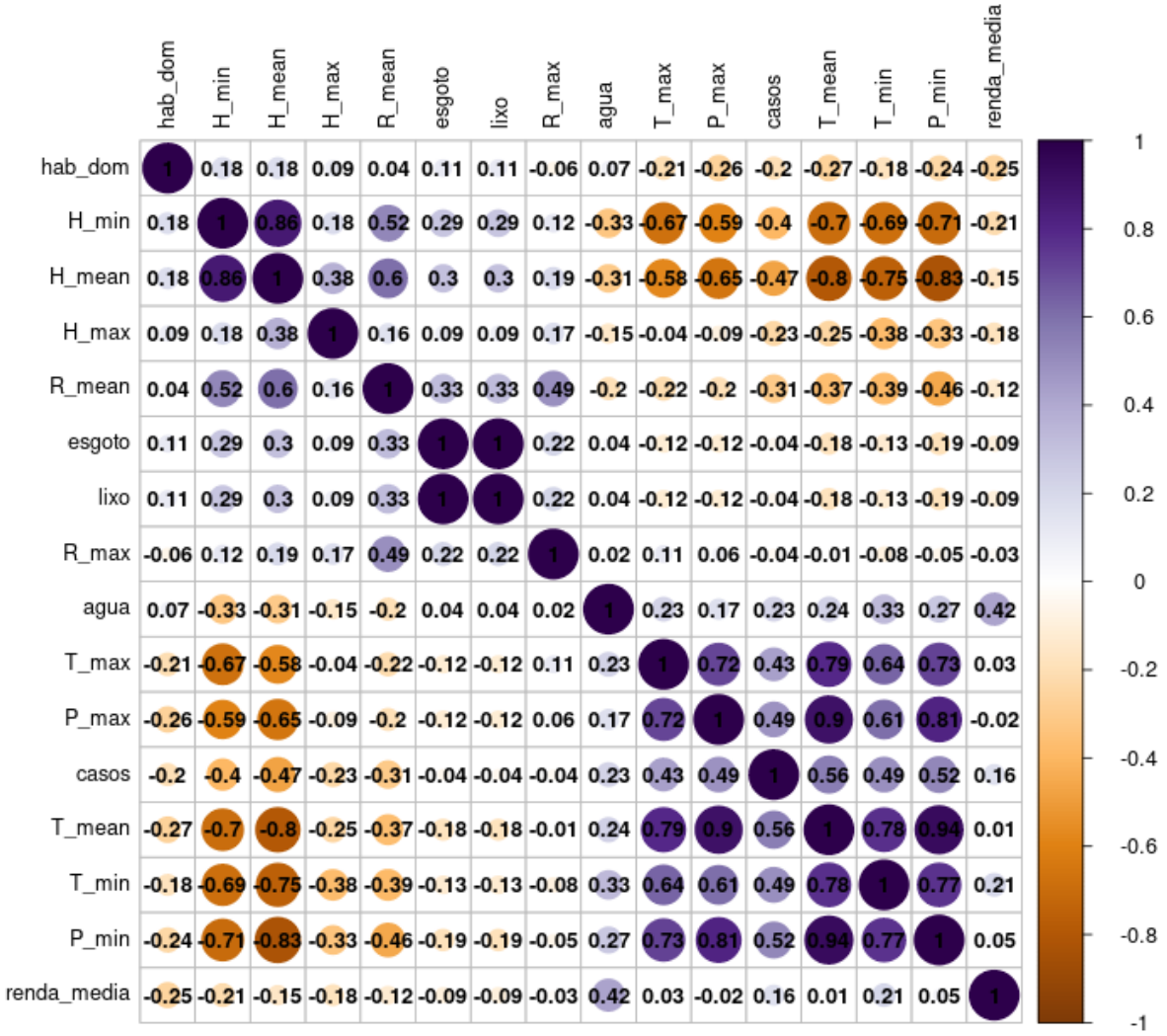

Figura 3: Correlação entre os dados climáticos, dados sociodemográficos e casos de incidência acumulada de dengue dos municípios do Estado de São Paulo no período de 01 de janeiro de 2015 a 31 de dezembro de 2019.

4.3 Análise de componentes principais

A Figura 4 mostra a porcentagem da variância explicada por cada componente principal. Podemos observar que a dimensão 1 e 2 conseguem explicar juntas 55,48% da variância dos dados. Utilizando o critério de Kaiser, verificou-se que as 15 dimensões iniciais podem ser reduzidas a cinco, as quais explicam 80,31% da variância dos dados originais.

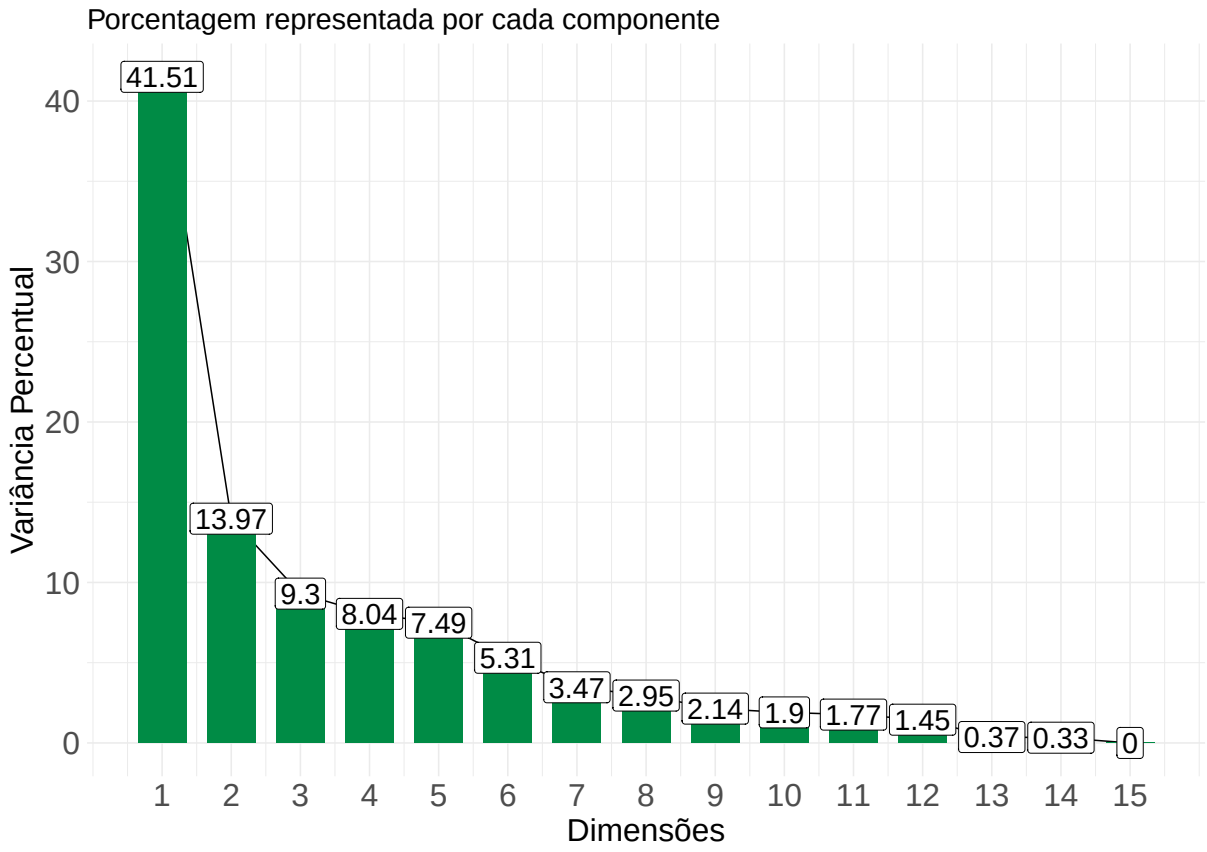

Figura 4: Porcentagem da variância explicada por cada componente principal.

A Figura 4 exibe as porcentagens da variância explicada por cada uma das componentes. Nota-se que a componente principal 1 (PC1) está mais associada às variáveis climáticas. No caso da temperatura média (T_{mean}), temperatura máxima (T_{max}), temperatura mínima (T_{min}), índice P máximo (P_{max}) e índice P mínimo, (P_{min}) a correlação é forte e diretamente proporcional. Em contrapartida, ao que ocorre com a umidade média (H_{mean}), pluviosidade média (R_{mean}) e umidade máxima (H_{max}) onde a correlação é inversamente proporcional. Já a dimensão 2 e 3 estão correlacionadas com as variáveis

sociodemográficas, a dimensão 2 com a quantidade de esgoto gerada por domicílios (esgoto) e a quantidade de lixo gerada por domicílios (lixo) e a dimensão 3 com a renda média por domicílios (renda média) e a quantidade de água consumida por domicílios (água) onde para ambas as correlações são positivas. A dimensão 4 possui correlação com a pluviosidade máxima (R_{max}) e a dimensão 5 com a variável que representa a quantidade de habitantes por domicílio (hab_{dom}).

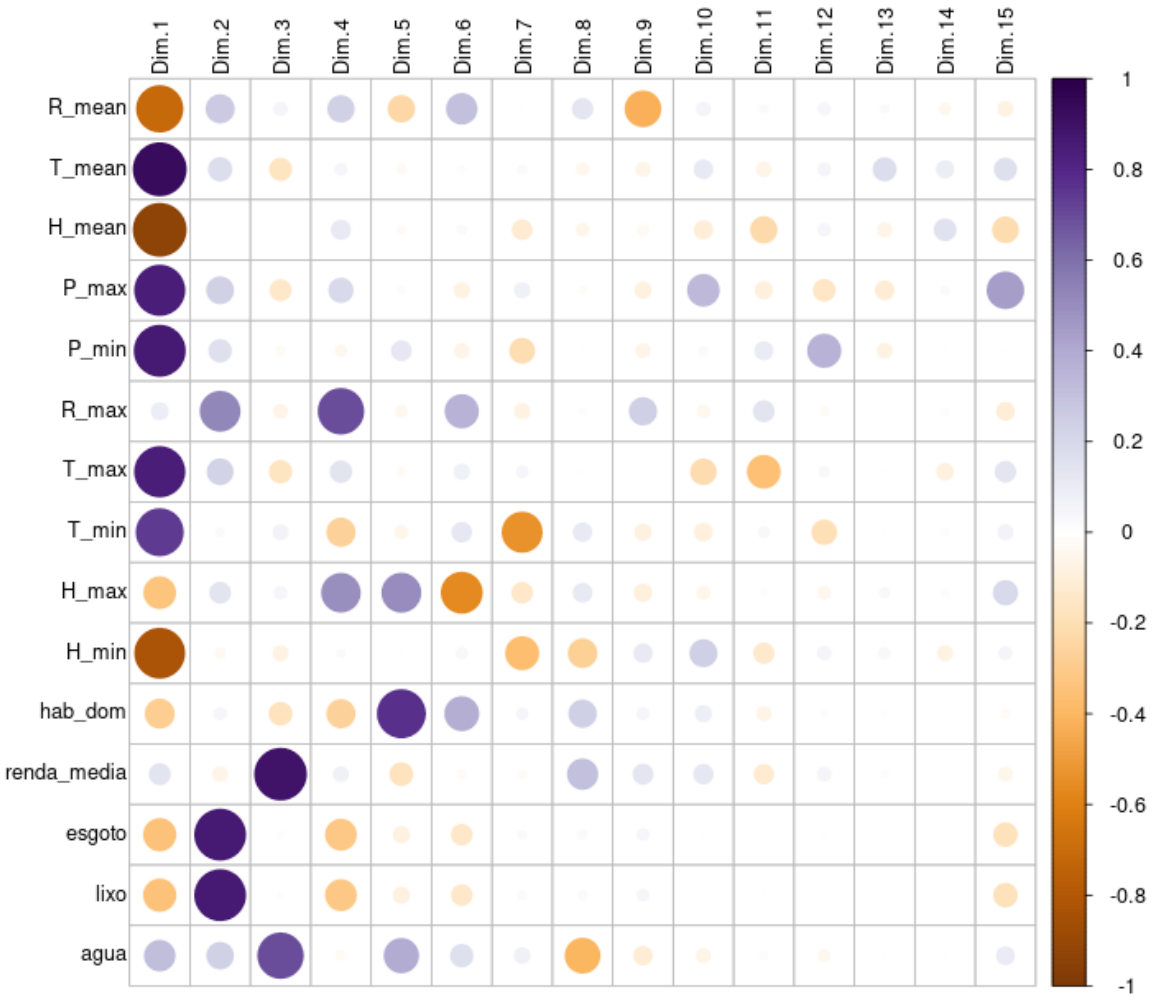

Figura 5: Correlação entre as componentes principais calculadas e as variáveis estudadas.

4.4 Análise de agrupamento

Utilizando as cinco primeiras componentes principais selecionadas utilizando o critério de Kaiser, foi realizado um agrupamento não hierárquico, através da técnica de *K-means*.

Utilizando o método do cotovelo, o qual testa várias quantidades diferentes de grupos e diz, qual representa o número ótimo de agrupamentos, foi selecionado a quantidade de agrupamentos a partir do qual o "ganho"para minimizar a soma dos quadrados da distância entre as observações e o centroide do grupo ao qual elas pertencem apresenta diminuição significativa (ver Figura 6). Por exemplo, podemos observar na Figura 6 que a inclinação da melhor reta ajustada no intervalo número de agrupamento entre 1 e 5 é maior que a obtida no intervalo entre 5 e 10. Portanto, para o problema proposto, cinco agrupamentos foram escolhidos, pois parece razoável.

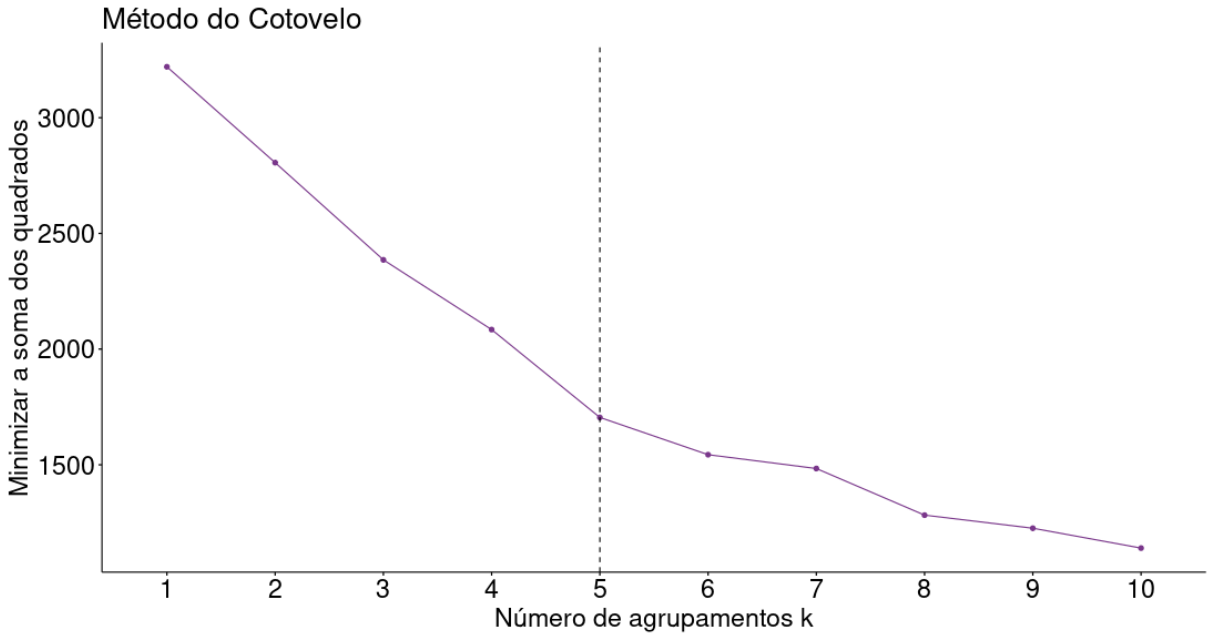

Figura 6: Soma dos quadrados da distância entre as observações e o centroide do grupo ao qual elas pertencem. O número de agrupamentos ideal é 5.

A Figura 7 mostra os grupos formados através do procedimento de K-médias. Nota-se que a representação apresenta sobreposição entre os grupos, resultado da projeção das cinco dimensões utilizadas nas duas dimensões representadas.

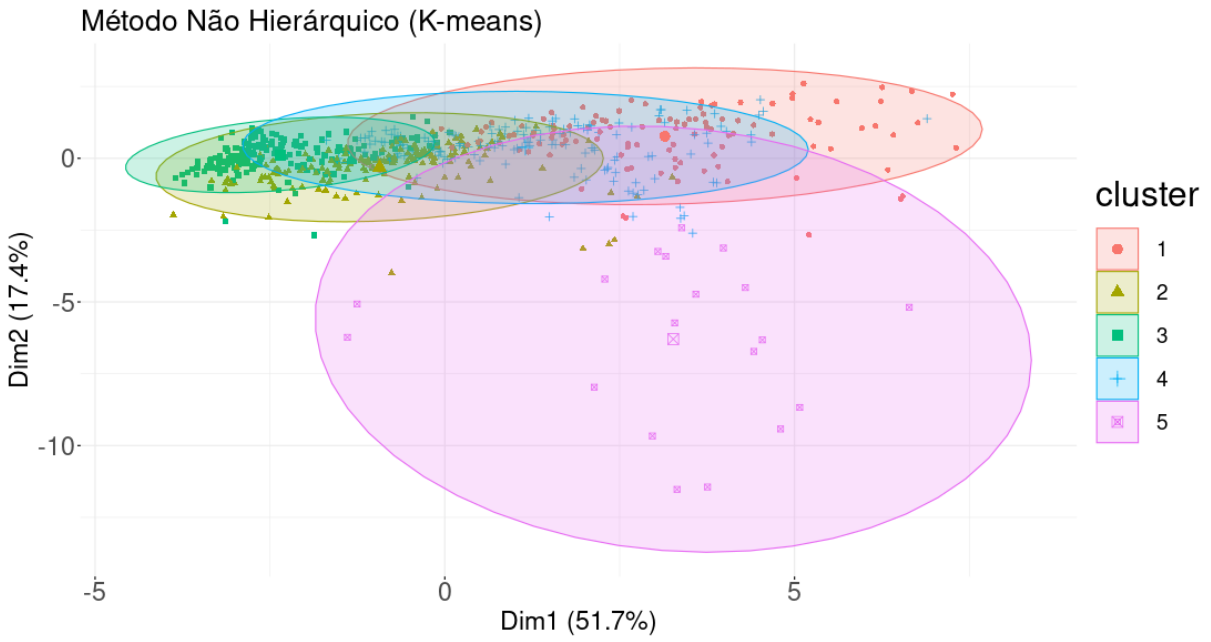

Figura 7: Agrupamento obtido pelo Método Não Hierárquico (*K-means*).

Logo após ser realizado o agrupamento não-hierárquico através do método *K-means*, foi realizado o agrupamento hierárquico. Com esta técnica, ao ser aplicada a distância euclidiana entre os dados e, através do coeficiente de correlação cofenético, que mede o grau de preservação das distâncias emparelhadas pelo dendograma em relação às distâncias originais, verificou-se que o método do encadeamento médio era o melhor a ser utilizado [7]. Em seguida, utilizou-se um corte para observar cinco grupos e comparar estes com os obtidos no método não-hierárquico (ver Figura 8).

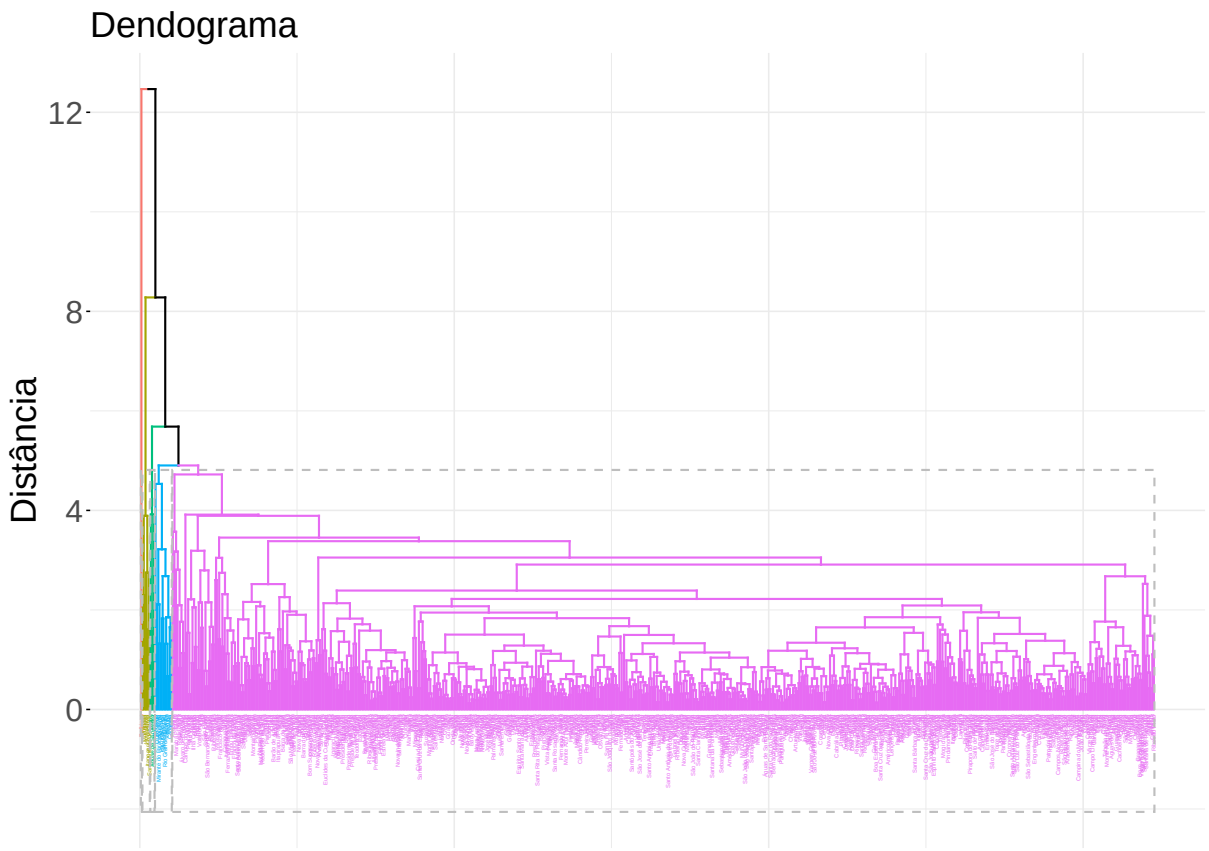

Figura 8: Dendrograma construído a partir da técnica de agrupamento hierárquica utilizando a distância euclidiana entre as observações e o método de ligação média.

Figura 9 mostra a comparação entre os dois métodos. O método não-hierárquico promove uma partição mais igualitária entre os grupos no quesito número de cidades em cada grupo.

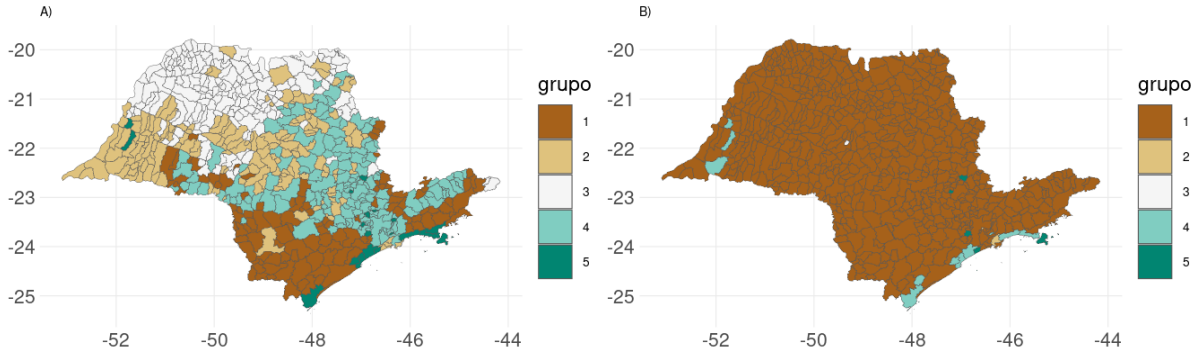

Figura 9: Mapa do Estado de São Paulo com os diferentes grupos encontrados através dos (a) método não-hierárquico e (b) método hierárquico.

4.5 Incidência de dengue nos diferentes grupos

A Figura 10 apresenta a comparação entre a incidência de dengue acumulada no período de 01 de janeiro de 2015 a 31 de dezembro de 2019 nos cinco grupos formados a partir do método não-hierárquico. Também traz os valores médios de algumas variáveis originais do estudo em cada grupo, são elas temperatura máxima, índice P mínimo, temperatura média, temperatura mínima, índice P máximo, umidade mínima, umidade média , pluviosidade média. Nota-se que o índice P e a temperatura têm comportamento semelhante ao observado nos dados de incidência de dengue, enquanto unidade e pluviosidade tem comportamento inverso.

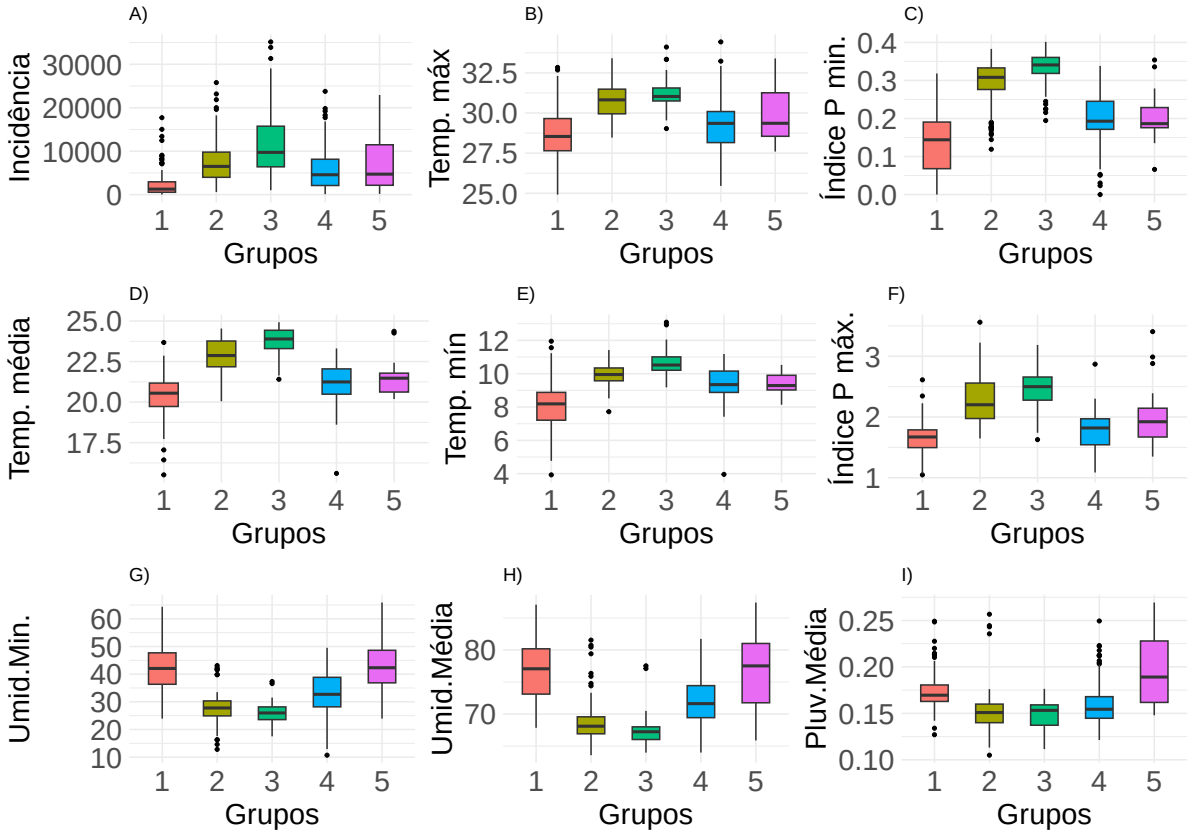

Figura 10: (A) Incidência de dengue acumulada por 100 mil habitantes, (B) Temperatura máxima, (C) Índice P mínimo, (D) Temperatura média, (E) Temperatura mínima, (F) Índice P máximo, (G) Umidade mínima, (H) Umidade média , (I) Pluviosidade média. Os dados são de cidades do Estado de São Paulo e compreendem o período de 2015 a 2019.

5 Conclusão

O estudo mostra que índice P, como mostrado pelos critérios que foram colocados, apresenta uma correlação moderada com a incidência acumulada de casos de dengue nas cidades do estado de São Paulo. As variáveis climáticas capturaram melhor o padrão de distribuição geográfica dos casos de dengue utilizadas nos agrupamentos ao contrário do que foi exibido para as sociodemográficas. Portanto, é possível utilizar os métodos aqui apresentados para a identificação de cidades alvo para a implementação de políticas públicas para o controle da dengue.

Bibliografia

1 NUNES, Priscila Conrado Guerra et al. 30 years of fatal dengue cases in Brazil: a review. **BMC public health**, BioMed Central, v. 19, n. 1, p. 1–11, 2019.

2 MINISTÉRIO DA SAÚDE. **São Paulo registrou 201 mil casos prováveis de dengue em 2023, entre janeiro e abril**. [S.l.: s.n.], mai. 2023. https://www.gov.br/saude/pt-br/assuntos/noticias-para-os-estados/sao-paulo/2023/maio/sao-paulo-registrou-201-mil-casos-provaveis-de-dengue-em-2023-e. (Acessado em 27/05/2023).

3 SINGHI, Sunit; KISSOON, Niranjan; BANSAL, Arun. Dengue e dengue hemorrágico: aspectos do manejo na unidade de terapia intensiva. **Jornal de Pediatria**, Sociedade Brasileira de Pediatria, v. 83, n. 2, s22–s35, mai. 2007. DOI: 10.1590/S0021-75572007000300004.

4 BRITO, Anderson Fernandes et al. Lying in wait: the resurgence of dengue virus after the Zika epidemic in Brazil. **Nature communications**, Nature Publishing Group UK London, v. 12, n. 1, p. 2619, 2021.

5 OBOLSKI, Uri et al. MVSE: An R-package that estimates a climate-driven mosquito-borne viral suitability index. **Methods in ecology and evolution**, Wiley Online Library, v. 10, n. 8, p. 1357–1370, 2019.

6 KRAEMER, Moritz UG et al. Big city, small world: density, contact rates, and transmission of dengue across Pakistan. **Journal of the Royal Society Interface**, The Royal Society, v. 12, n. 111, p. 20150468, 2015.

7 FÁVERO, Luiz Paulo; BELFIORE, Patrícia. **Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®**. [S.l.]: Elsevier Brasil, 2017.

8 ZAR, Jerrold H. **Biostatistical analysis**. [S.l.]: Pearson Education India, 1999.

9 QUINN, Gerald Peter; KEOUGH, Michael J. **Experimental design and data analysis for biologists**. [S.l.]: Cambridge university press, 2002.

10 HONGYU, Kuang; SANDANIELO, Vera Lúcia Martins;
OLIVEIRA JUNIOR, Gilmar Jorge de. Análise de componentes principais: resumo teórico, aplicação e interpretação. **E&S Engineering and science**, v. 5, n. 1, p. 83–90, 2016.

11 LINDEN, Ricardo. Técnicas de agrupamento. **Revista de Sistemas de Informação da FSMA**, n, v. 4, n. 4, p. 18–36, 2009.