UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS DO PONTAL

REBECA VIEIRA MACEDO MARQUES

Análise de agrupamentos aplicada a dados de criminalidade

Ituiutaba

2024

REBECA VIEIRA MACEDO MARQUES

Análise de agrupamentos aplicada a dados de criminalidade

Trabalho de Conclusão de Curso apresentado ao Instituto de Ciências Exatas e Naturais do Pontal da Universidade Federal de Uberlândia como requisito parcial para obtenção do título de bacharel em matemática.

Orientadora: Profa. Dra. Franciella Marques da Costa.

Ituiutaba

2024

REBECA VIEIRA MACEDO MARQUES

Análise de agrupamentos aplicada a dados de criminalidade

Trabalho de Conclusão de Curso apresentado ao Instituto de Ciências Exatas e Naturais do Pontal da Universidade Federal de Uberlândia como requisito parcial para obtenção do título de bacharel em matemática.

Ituiutaba, 25 de abril de 2024

Banca Examinadora:

_____

Orientadora: Profa. Dra. Franciella Marques da Costa – ICENP/UFU

_____

Prof. Dr. Carlos Eduardo Petronilho Boiago – ICENP/UFU

_____

Profa. Dra. Gabriella de Freitas Alves – ICENP/UFU

AGRADECIMENTOS

Primeiramente, expresso minha profunda gratidão a Deus pela vida que me foi concedida e por sua constante presença, guiando-me e fortalecendo-me em todos os momentos.

Também quero dedicar palavras de apreço e admiração à minha mãe, Glaudilene. Desde os meus dois anos de idade, ela assumiu a árdua tarefa de cuidar de mim e da minha irmã, após o falecimento de meu pai. Sua dedicação incansável, sua sabedoria na educação que nos proporcionou e a amizade que sempre nos dispensou são tesouros inestimáveis em minha vida. Ela é a minha base, meu exemplo e meu alicerce.

Não posso deixar de reconhecer e agradecer a minha irmã, Ester. Além de ser minha irmã, ela é minha amiga mais próxima, uma aliada incansável que esteve ao meu lado em cada desafio, incentivando-me e compartilhando comigo alegrias e tristezas.

Agradeço de coração à minha avó, Jailta, que é a matriarca de nossa família. Seus ensinamentos sábios e seus preciosos conselhos moldaram não apenas a mim, mas a todas nós, quatro mulheres, e a toda a nossa família. Sua presença é uma bênção que ilumina nossas vidas.

Também expresso minha gratidão ao meu esposo, Lucas, e a toda a minha família, por seu apoio inabalável e suas palavras de conforto nos momentos mais desafiadores. Sua presença é um farol de amor e apoio em meu caminho.

Não poderia deixar de mencionar a importância dos professores e da instituição em minha jornada. Inúmeras vezes, eles me guiaram, inspiraram e auxiliaram em meu crescimento acadêmico e pessoal. Em especial, agradeço à minha orientadora, Franciella, cuja orientação e incentivo foram fundamentais para o meu desenvolvimento.

Com gratidão e apreço, dedico este trabalho a todos que moldaram quem sou e contribuíram para minhas conquistas em especial à minha família.

RESUMO

A análise de agrupamentos é uma importante ferramenta da multivariada que pode ser utilizada em diversas áreas, inclusive em dados de criminalidade. O objetivo deste trabalho foi analisar dados de criminalidade, dos 26 estados brasileiros e do distrito federal, utilizando análise de agrupamentos. Para a análise de agrupamentos utilizou-se diferentes métodos e diferentes distâncias, em seguida obteve-se a correlação cofenética para definir qual combinação resultou em um melhor agrupamento e utilizou-se o pacote NbClust, do software R, para definir o número de clusters na partição final. A análise de agrupamentos para os dados de criminalidade, nas 27 unidades federativas do Brasil para o ano de 2021, mostrou-se eficiente, formando 4 grupos com correlação cofenética igual a 0,88.

Palavras-chave: análise de agrupamentos; correlação cofenética; criminalidade; estatística multivariada; técnicas hierárquicas aglomerativas.

ABSTRACT

Cluster analysis is an important multivariate tool that can be used in various areas, including crime data. The aim of this study was to analyze crime data from the 26 Brazilian states and from the federal district using cluster analysis. Different methods and different distances were used for the cluster analysis, then the cophenetic correlation was obtained to define which combination resulted in the best cluster and the NbClust package in the R software was used to define the number of clusters in the final partition. The cluster analysis for crime data in Brazil's 27 federal units for the year 2021 proved to be efficient, forming 4 groups with a cophenetic correlation of 0.88.

Keywords: cluster analysis; cophenetic correlation; crime; multivariate statistics; hierarchical agglomerative techniques.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

A estatística vem sendo utilizada de forma significativa na tomada de decisões em diversas áreas do conhecimento, tais como: economia, administração, contabilidade, medicina entre outras. Dificilmente, um pesquisador ao conduzir um estudo está interessado em apenas uma variável. Dessa forma, a estatística multivariada é uma área da estatística de extrema importância.

As técnicas da estatística multivariada consistem em métodos para análise simultânea de um conjunto de variáveis que foram medidas em cada elemento amostral. Segundo Mingoti (2013), a estatística multivariada possui métodos exploratórios para simplificação da estrutura de variabilidade dos dados e técnicas para inferência estatística. A análise de componentes principais, análise fatorial, análise de correlações canônicas, análise de agrupamento, análise discriminante e análise de correspondência são técnicas exploratórias para simplificação da estrutura de variabilidade dos dados e os métodos de estimação dos parâmetros, testes de hipóteses, análise de variância, análise de covariância e análise de regressão multivariada são técnicas destinadas a inferência estatística (MINGOTI, 2013).

A análise de agrupamentos, também chamada de análise de conglomerado ou análise de cluster é uma importante ferramenta da multivariada que pode ser utilizada em diversas áreas, inclusive em dados de criminalidade. O estudo de dados de criminalidade é de extrema importância para a sociedade, podendo ser utilizado no direcionamento de medidas de segurança.

Ventorim e Netto (2023) analisou dados criminais da cidade do Rio de Janeiro, De Figueiredo, et al. (2020) utilizou dados do estado do Espírito Santo com o objetivo de identificar fatores determinantes para o controle dos índices de criminalidade, Guimarães e Becker (2021) analisaram a distribuição espacial das taxas de roubos, tráfico de entorpecentes e os homicídios no Rio Grande do Sul e Nascimento (2019) realizou uma análise de agrupamento aplicada a dados de violência no Brasil.

Diante do exposto, o objetivo deste trabalho foi analisar dados de criminalidade, dos 26 estados brasileiros e do distrito federal, utilizando análise de agrupamento, que é uma técnica da estatística multivariada. Este estudo pode ser utilizado para auxiliar a tomada de decisões na área de segurança pública.

2. REFERENCIAL TEÓRICO

2.1 Criminalidade no Brasil

De acordo com o Fórum Brasileiro de Segurança Pública – FBSP (2022), em 2020 o Brasil tinha 2,7% dos habitantes do planeta e contabilizava 20,5% dos homicídios conhecidos que ocorreram. Em 2020, ocorreram 232.676 assassinatos em um total de 102 países, enquanto somente no Brasil ocorreram 50.512 (FBSP, 2022).

Segundo o FBSP (2023), o Brasil é um país violento. Em 2022 houve 47.452 Mortes Violentas Intencionais (MVI). As mortes violentas intencionais é uma categoria que inclui vítimas de homicídio doloso (incluindo feminicídios e policiais assassinados), roubos seguidos de morte, lesão corporal seguida de morte e as mortes provenientes de intervenções policiais (FBSP, 2023).

Conforme o FBSP (2023), em 2022 o Amapá foi o estado com maior taxa de MVI do país, sendo uma taxa de 50,6 por 100 mil habitantes. A Bahia ocupa a segunda posição com uma taxa de 47,1 por 100 mil habitantes e na terceira posição tem-se o Amazonas com taxa de 38,8 por 100 mil habitantes. São Paulo apresentou uma taxa de MVI de 8,4 por 100 mil habitantes, sendo o estado com menor taxa de mortes violentas intencionais, Santa Catarina tem 9,1 por 100 mil habitantes e o Distrito Federal possui uma taxa de 11,3 por 100 mil habitantes. Observando todos os estados brasileiros, 20 estados registraram taxas de mortes violentas intencionais maiores que a média nacional.

2.2 Análise de agrupamentos

A análise de agrupamentos, também chamada de análise de conglomerado ou análise de cluster é uma das técnicas da estatística multivariada que pode ser utilizada em diversas situações. O objetivo da análise de agrupamentos é separar os elementos da população ou da amostra em grupos, nos quais os elementos pertencentes ao mesmo grupo sejam homogêneos entre si considerando as variáveis que foram observadas e os elementos que estão em grupos diferentes sejam heterogêneos, considerando as mesmas variáveis (MINGOTI, 2013).

Segundo Mingoti (2013) e Ferreira (2008), as técnicas de agrupamentos podem ser divididas em técnicas hierárquicas e não hierárquicas. As técnicas de agrupamentos

hierárquicas podem ser classificadas em agrupamento hierárquico aglomerativo ou agrupamento hierárquico divisivo. No agrupamento hierárquico aglomerativo começamos o processo com n grupos, em que n é o número de elementos pertencentes a amostra, ou seja, cada grupo é formado por um único elemento. A cada passo vão sendo formados novos grupos até que se forme um único grupo com todos os elementos. Já no agrupamento hierárquico divisivo, temos o processo inverso, começamos o processo com um único grupo formado por n elementos e a cada passo vão se formando novos grupos até o momento em que se obtenha n grupos com um elemento cada.

Para se realizar uma análise de agrupamento é necessário utilizar uma medida de similaridade ou de dissimilaridade para que seja possível identificar os elementos que são mais semelhantes ou não.

2.3 Distâncias

Para realizar uma análise de agrupamentos, inicialmente é necessário escolher a medida de similaridade ou de dissimilaridade que será utilizada. Existem várias medidas de similaridade e de dissimilaridade e cada uma dessas medidas pode gerar agrupamentos diferentes. De acordo com Mingoti (2013), as medidas de dissimilaridade são indicadas para variáveis quantitativas e quanto menor o valor obtido maior a similaridade entre os elementos comparados. As medidas de distâncias, apresentadas a seguir, são medidas de dissimilaridade:

- **Distância Euclidiana**

Segundo Fávero, et al. (2009), a distância euclidiana entre duas observações, representadas por i e j, é calculada como a raiz quadrada da soma dos quadrados das diferenças entre cada par de observações correspondentes a i e j, considerando todas as p variáveis envolvidas no estudo. A distância euclidiana é definida por:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

em que o valor x_{ik} é referente a variável k medida no elemento i e x_{jk} é o valor da variável k medida no elemento j.

- **Distância de Minkowski**

De acordo com Fávero, et al. (2009), a distância de Minkowski entre as observações i e j, é definida por:

$$d_{ij} = \left(\sum_{k=1}^{p}\left(\left|x_{ik} - x_{jk}\right|\right)^{n}\right)^{1/n}$$

em que p é o número de variáveis e $n = 1, 2, \cdots, \infty$. Diferentes valores para n resultam em diferentes distâncias. Por exemplo, para $n = 1$ tem-se a distância conhecida como distância de Manhattan e para $n = 2$, tem-se a distância euclidiana.

- **Distância de Manhattan**

De acordo com Fávero, et al. (2009), a distância de Manhattan entre as observações i e j, é definida por:

$$d_{ij} = \sum_{k=1}^{p}\left|x_{ik} - x_{jk}\right|$$

A distância de Manhattan também é conhecida por distância absoluta, bloco ou city-block.

- **Distância de Mahalanobis**

Segundo Fávero, et al. (2009), a distância de Mahalanobis é uma medida utilizada para avaliar a dissimilaridade entre dois elementos i e j, levando em consideração a estrutura de covariância dos dados. A distância de Mahalanobis é definida por:

$$d_{ij} = \sqrt{(x_i - x_j)' S^{-1}(x_i - x_j)}$$

em que S^{-1} é a matriz inversa da matriz de covariâncias amostrais.

2.4 Métodos hierárquicos aglomerativos

No método aglomerativo, inicialmente, cada elemento é tratado como um grupo independente. À medida que o processo avança, grupos ou elementos são fundidos com base em sua semelhança, resultando em um último estágio onde um único grupo é formado, incorporando todos os elementos (MINGOTI, 2013).

De acordo com Fávero, et al. (2009), assim que for formado o primeiro cluster é necessário definir como a distância entre dois clusters será calculada. Existem vários métodos disponíveis para formar os agrupamentos, sendo que a principal diferença entre eles é a maneira como são calculadas as distâncias entre os grupos já formados e os que ainda restam para ser agrupados. A seguir serão descritos os métodos de agrupamentos hierárquicos mais utilizados.

- **Método da ligação individual ou menor distância (Single Linkage ou Nearest Neighbor)**

De acordo com Fávero, et al. (2009), no método de ligação individual o primeiro grupo será formado pelo vizinho mais próximo, ou seja, será formado pelos dois elementos que tiverem a menor distância entre eles. No passo seguinte, o elemento que possuir a menor distância em relação a este grupo, passa a fazer parte do grupo. Este processo é repetido até a formação de um único grupo.

Considere dois grupos (i e j) e (k), a distância entre eles é definida pela distância mínima entre um ponto qualquer de um grupo até um ponto qualquer do outro grupo.

A distância entre os grupos será definida como:

$$d_{(ij)(k)} = \min\{d_{ik}, d_{jk}\}$$

- **Método da maior distância ou ligação completa (Complete Linkage ou Furthest Neighbor)**

De acordo com Fávero, et al. (2009), no método de ligação completa a distância entre dois grupos é obtida considerando a distância máxima entre todos os pares de observações dos dois grupos. São agrupados em um mesmo grupo aqueles que possuírem o menor valor de máximo.

Considere dois grupos (i e j) e (k), a distância entre eles é definida pela distância máxima entre um ponto qualquer de um grupo até um ponto qualquer do outro grupo.

A distância entre os grupos será definida como:

$$d_{(ij)(k)} = \max\{d_{ik}, d_{jk}\}$$

- **Método da distância média ou ligação média (Average Linkage ou Between Groups)**

Segundo Fávero, et al. (2009), no método de ligação média a distância entre dois grupos é obtida considerando a distância média entre os pares de elementos dos dois grupos, com o objetivo de agrupar os que possuem a menor distância média.

A distância entre os grupos será definida como:

$$d_{(ij)(k)} = \text{média}\{d_{ik}, d_{jk}\}$$

- **Método do Centróide**

Segundo Mingoti (2013), no método do centróide a distância entre dois grupos é determinada pela distância euclidiana ao quadrado ou a distância euclidiana entre os vetores de médias, chamados de centróides, dos grupos em análise.

Apesar da simplicidade do método do centróide, em cada fase é necessário recorrer aos dados originais para calcular a matriz de distâncias, resultando em um tempo computacional mais prolongado em comparação com outras abordagens. Ademais, ao contrário de outros métodos, o método do centróide não pode ser utilizado em situações em que apenas a matriz de distância entre os elementos amostrais está disponível. (MINGOTI, 2013).

- **Método de Ward**

Segundo Reis (2001) apud Fávero, et al. (2009), o método de Ward pode ser descrito nos seguintes passos:

- Calcula-se as médias das variáveis para cada grupo;

- Calcula-se o quadrado da distância euclidiana entre as médias obtidas no passo anterior e os valores das variáveis para cada elemento;

- Em seguida, obtém-se a soma das distâncias para todos os elementos;

- O objetivo é minimizar a variância dentro dos grupos. Dessa forma, no último passo, deve-se minimizar a função chamada soma dos quadrados dos erros ou soma dos quadrados dentro dos grupos.

Após a realização do agrupamento é possível construir um gráfico chamado de dendrograma ou dendograma. O dendrograma é uma representação gráfica em forma de árvore em que é possível visualizar a história do agrupamento (MINGOTI, 2013).

2.5 Correlação cofenética

Segundo Da Silva (2016), a utilização de diferentes distâncias e diferentes métodos de agrupamento podem gerar diferentes agrupamentos. Após a obtenção do agrupamento é importante avaliar a qualidade do resultado, ou seja, o quanto as distâncias originais se assemelham às distâncias geradas pelo dendrograma. Para realizar essa avaliação pode-se utilizar o coeficiente de correlação cofenética, em que quanto maior a correlação cofenética, menor será a distorção causada pelo agrupamento.

Segundo Da Silva e Dias (2013), o coeficiente de correlação cofenética é definido por:

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j>i}^{n}\left(c_{ij}-\bar{c}\right)\left(d_{ij}-\bar{d}\right)}{\left(\sum_{i=1}^{n-1}\sum_{j>i}^{n}\left(c_{ij}-\bar{c}\right)^2\right)^{1/2}\left(\sum_{i=1}^{n-1}\sum_{j>i}^{n}\left(d_{ij}-\bar{d}\right)^2\right)^{1/2}}$$

em que, $\bar{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j>i}^{n} c_{ij}$, $\bar{d} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j>i}^{n} d_{ij}$, c_{ij} é o elemento da linha i e coluna j da matriz de distâncias produzida pelo dendrograma, d_{ij} é o elemento da linha i e coluna j da matriz de distâncias originais e n é o número de elementos da amostra.

Saraçli e Akşit (2022) realizou a análise de agrupamento utilizando diferentes métodos e diferentes distâncias, depois utilizou a correlação cofenética para definir qual combinação resultou em um melhor agrupamento.

2.6 Número de cluster para a partição final

Existem diversos critérios para determinar o número ideal de clusters na partição final. Mingoti (2013), por exemplo, cita sete critérios que podem ser utilizados para definir o número de clusters. Um desses critérios é o corte do dendrograma, que envolve

a observação de pontos de salto significativos em relação aos demais, indicando uma perda acentuada de similaridade entre os grupos formados.

Charrad, et al. (2014) criaram o pacote NbClust, disponível no software R. Este pacote possui uma função que calcula 30 índices indicados para a obtenção do número ideal de clusters. A função retorna o número ideal de clusters baseado na quantidade indicada pela maioria dos índices.

3. MATERIAIS E MÉTODOS

Os dados utilizados neste trabalho são referentes aos dados de criminalidade das 27 unidades federativas do Brasil no ano de 2021. O conjunto de dados foi obtido a partir do website do Fórum Brasileiro de Segurança Pública (FBSP, 2024).

Para realizar a análise de agrupamentos foram consideradas as taxas, por 100 mil habitantes, das seguintes variáveis:

- X1 – homicídio doloso em número de vítimas.
- X2 – latrocínio.
- X3 – lesão corporal seguida de morte.
- X4 – estupro (incluindo estupro de vulnerável).
- X5 – roubo e furto de veículos.
- X6 – feminicídio.
- X7 – mortes decorrentes de intervenções militares.

Inicialmente foi realizada a análise descritiva das variáveis em estudo. Foi apresentado o gráfico de barras, o valor mínimo, o valor máximo, a média, o primeiro quartil, a mediana, o terceiro quartil e o desvio padrão para cada variável em estudo.

Para realizar a análise de agrupamentos, as variáveis foram padronizadas. Em seguida utilizou-se diferentes distâncias e métodos de agrupamentos. Foram utilizadas a distância euclidiana, a distância de Manhattan e a distância de Mahalanobis. Foram utilizados o método da ligação individual, o método da ligação completa, o método da ligação média, o método do centróide e método de Ward.

Para avaliar a qualidade do agrupamento utilizou-se a correlação cofenética. Foi escolhido a combinação de distância e método com maior correlação cofenética. O resultado do agrupamento foi apresentado no dendrograma.

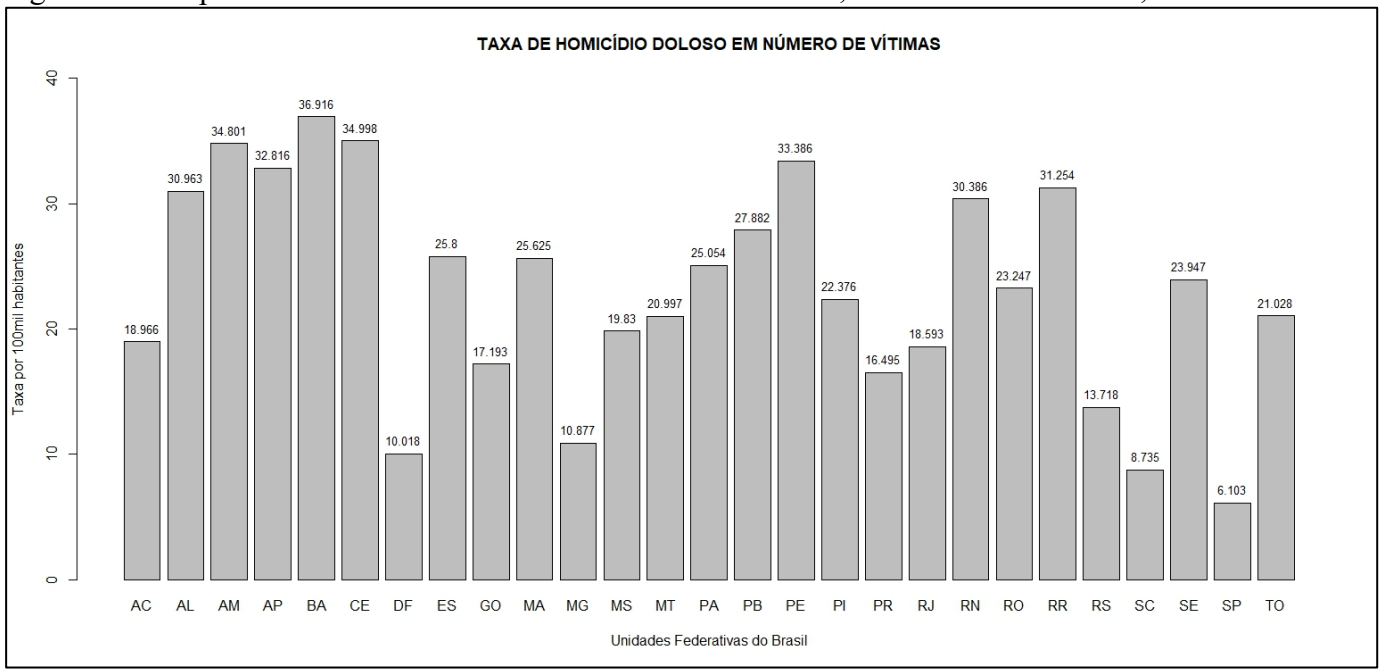
Para definir o número de cluster a ser considerado, foi utilizado o pacote NbClust do software R. Foi escolhido o número de clusters definido pela maioria dos índices.

Todas as análises estatísticas foram realizadas utilizando o software R (R CORE TEAM, 2024). Utilizou-se o pacote biotools (DA SILVA; MALAFAIA; MENEZES, 2017) (DA SILVA, 2021) para o cálculo da distância de Mahalanobis, o pacote NbClust (CHARRAD et al., 2014) foi utilizado para definir a número de clusters na partição final e o pacote factoextra (KASSAMBARA; MUNDT, 2020) foi utilizado para obter o dendrograma.

4. RESULTADOS E DISCUSSÕES

Inicialmente foi obtido os gráficos de barras para todas as variáveis em estudo. Na figura 1 é apresentado o gráfico referente ao crime homicídio doloso em números de vítimas (X1). Observa-se que a maior ocorrência desse tipo de crime foi na Bahia com uma taxa de 36,916 a cada 100 mil habitantes seguido por Ceará com 34,998 e São Paulo foi o estado com menor taxa de ocorrência, sendo 6,103 a cada 100 mil habitantes.

Figura 1: Taxa por 100 mil habitantes do crime homicídio doloso, em número de vítimas, nas 27 unidades federativas do Brasil no ano de 2021.[1]

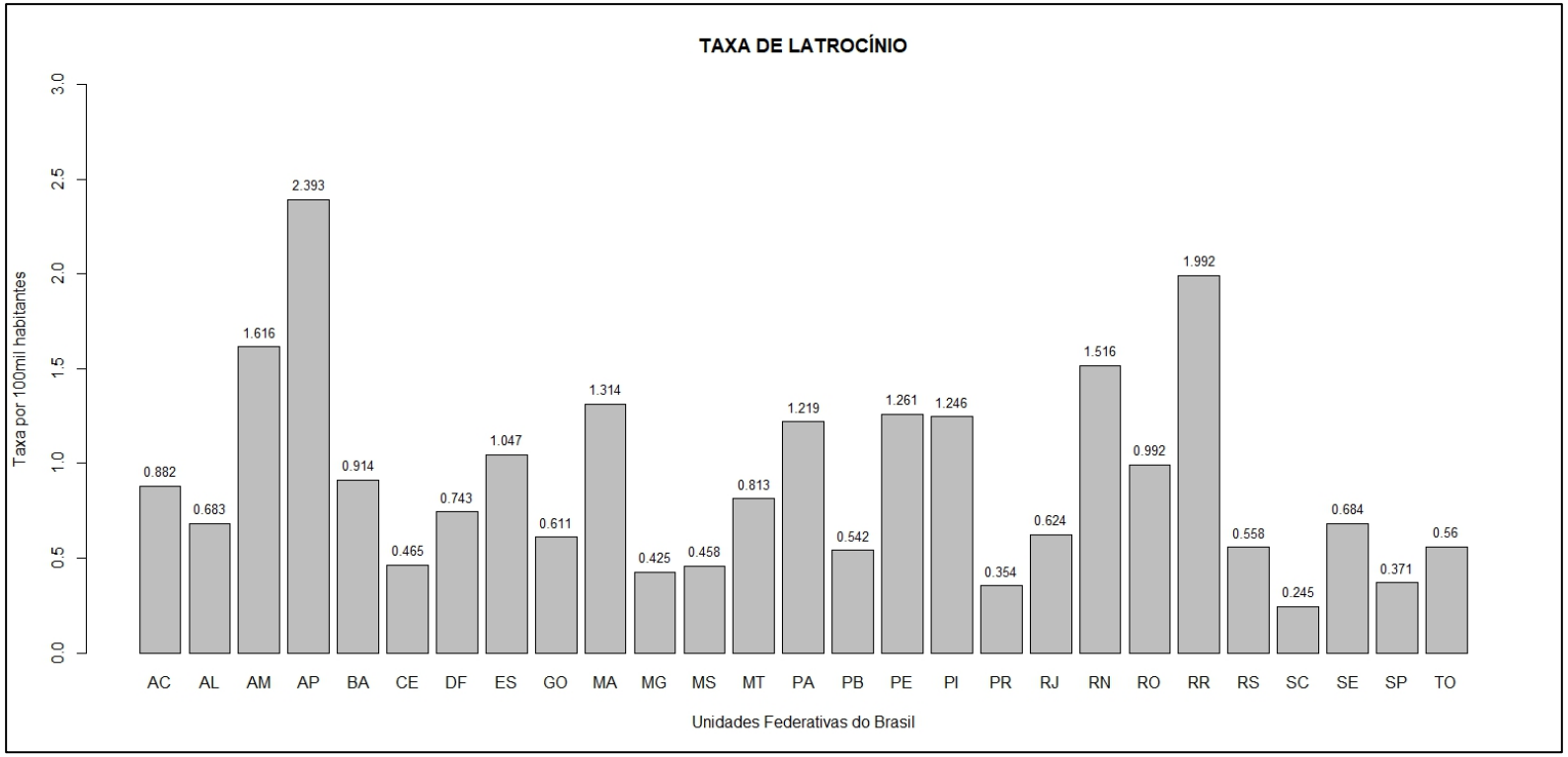


Fonte: A autora.

[1]AC – Acre, AL – Alagoas, AM – Amazonas, AP – Amapá, BA – Bahia, CE – Ceará, DF – Distrito Federal, ES – Espírito Santo, GO – Goiás, MA – Maranhã, MG – Minas Gerais, MS – Mato Grosso do Sul, MT – Mato Grosso, PA – Pará, PB – Paraíba, PE – Pernambuco, PI – Piauí, PR – Paraná, RJ – Rio de Janeiro, RN – Rio Grande do Norte, RO – Rondônia, RR – Roraima, RS – Rio Grande do Sul, SC – Santa Catarina, SE – Sergipe, TO – Tocantins.

Na figura 2 tem-se o gráfico referente ao crime latrocínio (X2). Pode-se observar no gráfico que o crime latrocínio acontece em menor quantidade se comparado com o homicídio doloso. Esse tipo de crime ocorreu com maior incidência no Amapá com uma taxa de 2,393 a cada 100 mil habitantes seguido por Roraima com 1,992 e com menor ocorrência tem-se o estado de Santa Catarina com taxa de 0,245.

Figura 2: Taxa por 100 mil habitantes do crime latrocínio, nas 27 unidades federativas do Brasil no ano de 2021.[2]

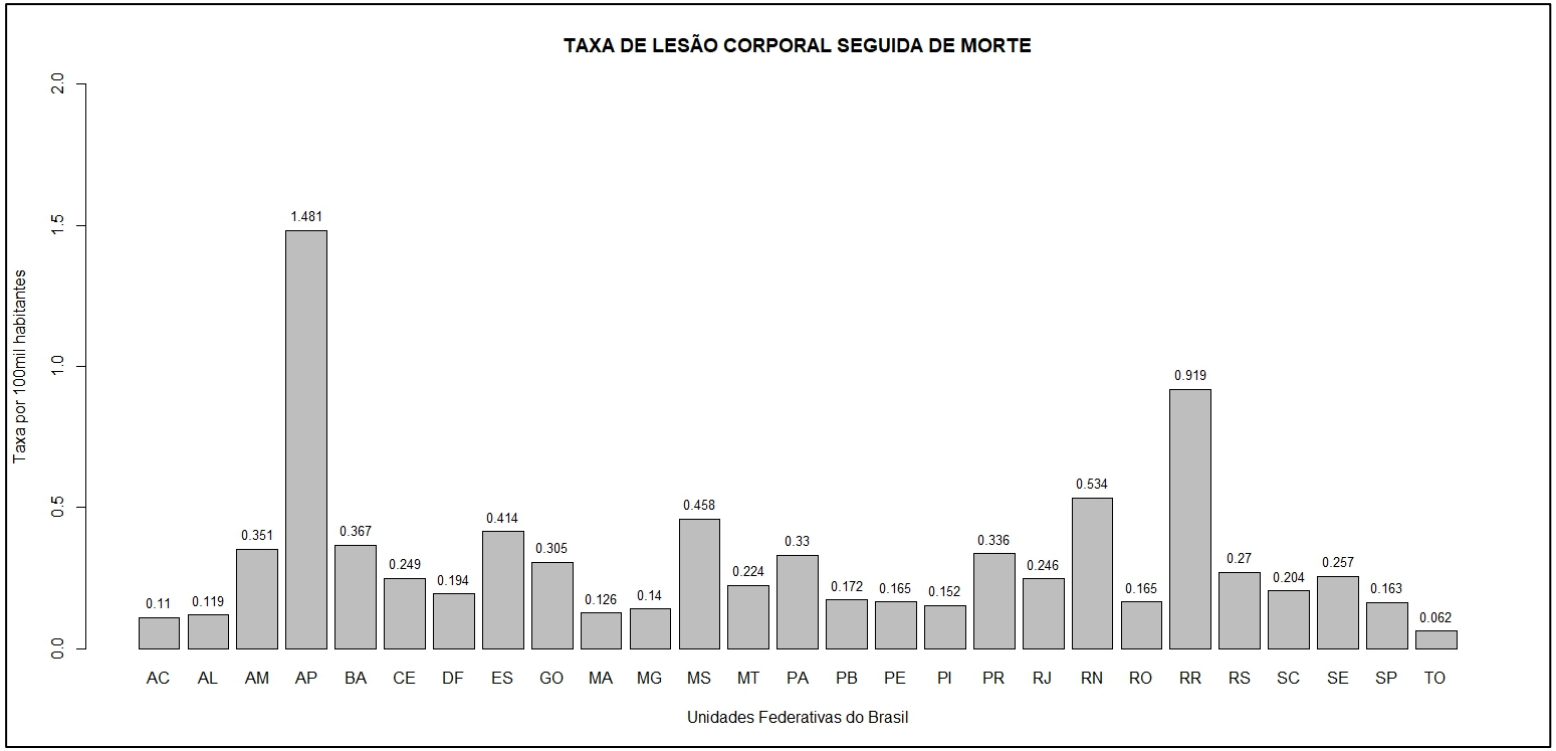


Fonte: A autora.

[2] AC – Acre, AL – Alagoas, AM – Amazonas, AP – Amapá, BA – Bahia, CE – Ceará, DF – Distrito Federal, ES – Espírito Santo, GO – Goiás, MA – Maranhã, MG – Minas Gerais, MS – Mato Grosso do Sul, MT – Mato Grosso, PA – Pará, PB – Paraíba, PE – Pernambuco, PI – Piauí, PR – Paraná, RJ – Rio de Janeiro, RN – Rio Grande do Norte, RO – Rondônia, RR – Roraima, RS – Rio Grande do Sul, SC – Santa Catarina, SE – Sergipe, TO – Tocantins.

Na figura 3 tem-se o gráfico referente ao crime lesão corporal seguida de morte (X3). Observa-se que a variável X3 ocorre com maior incidência no Amapá com uma taxa de 1,481 a cada 100 mil habitantes seguido por Roraima com 0,919 e com menor ocorrência o estado do Tocantins com uma taxa de 0,062 a cada 100 mil habitantes.

Figura 3: Taxa por 100 mil habitantes do crime lesão corporal seguida de morte, nas 27 unidades federativas do Brasil no ano de 2021.[3]

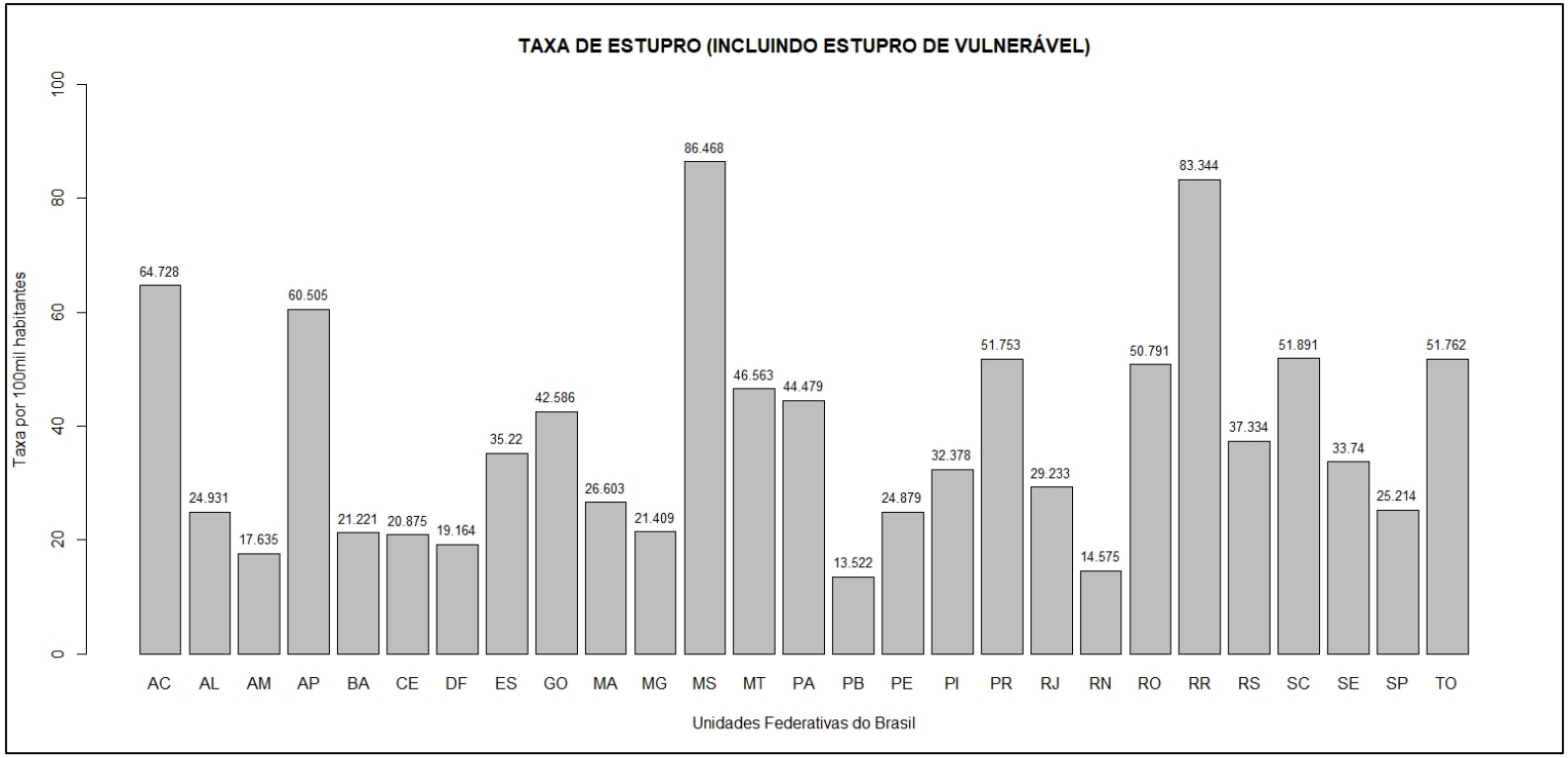


Fonte: A autora.

[3]AC – Acre, AL – Alagoas, AM – Amazonas, AP – Amapá, BA – Bahia, CE – Ceará, DF – Distrito Federal, ES – Espírito Santo, GO – Goiás, MA – Maranhã, MG – Minas Gerais, MS – Mato Grosso do Sul, MT – Mato Grosso, PA – Pará, PB – Paraíba, PE – Pernambuco, PI – Piauí, PR – Paraná, RJ – Rio de Janeiro, RN – Rio Grande do Norte, RO – Rondônia, RR – Roraima, RS – Rio Grande do Sul, SC – Santa Catarina, SE – Sergipe, TO – Tocantins.

Observa-se na figura 4 o gráfico que representa o crime de estupro (incluindo estupro de vulnerável) (X4). O estado com maior taxa foi o Mato Grosso do Sul com 86,468 a cada 100 mil habitantes seguido por Roraima com 83,344 e com menor ocorrência tem-se o estado da Paraíba com 13,522 a cada 100 mil habitantes.

Figura 4: Taxa por 100 mil habitantes, do crime estupro (incluindo estupro de vulnerável), nas 27 unidades federativas do Brasil no ano de 2021.[4]

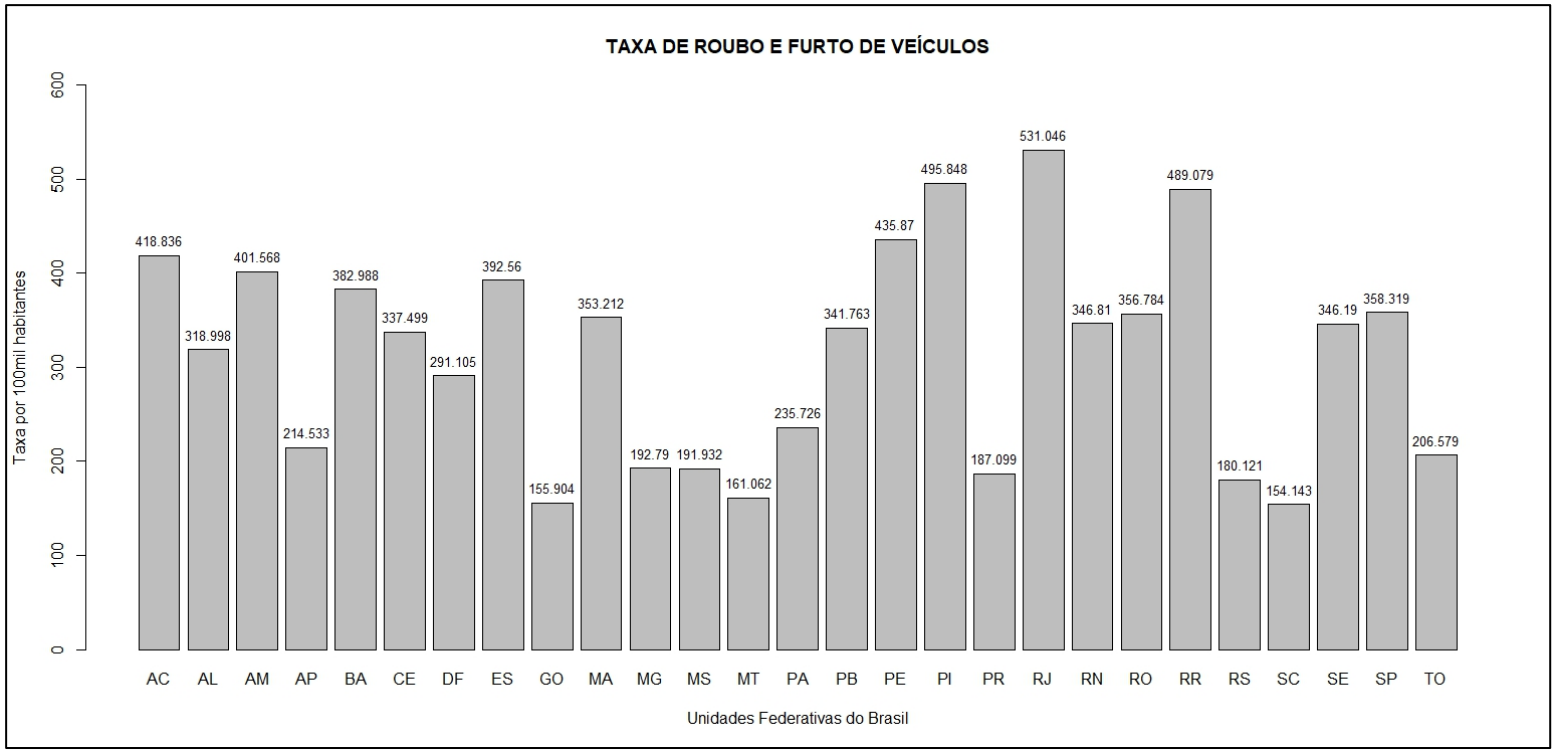


Fonte: A autora.

[4] AC – Acre, AL – Alagoas, AM – Amazonas, AP – Amapá, BA – Bahia, CE – Ceará, DF – Distrito Federal, ES – Espírito Santo, GO – Goiás, MA – Maranhã, MG – Minas Gerais, MS – Mato Grosso do Sul, MT – Mato Grosso, PA – Pará, PB – Paraíba, PE – Pernambuco, PI – Piauí, PR – Paraná, RJ – Rio de Janeiro, RN – Rio Grande do Norte, RO – Rondônia, RR – Roraima, RS – Rio Grande do Sul, SC – Santa Catarina, SE – Sergipe, TO – Tocantins.

Na figura 5 é mostrado o gráfico referente ao crime roubo e furto de veículos (X5), em que é possível observar que esse crime ocorreu com uma taxa bem maior que os crimes abordados anteriormente. A maior incidência registrada é no Rio de Janeiro com uma taxa de 531,046 a cada 100 mil habitantes seguido pelo Piauí com 495,848 e com menor ocorrência tem-se o estado de Santa Catarina com uma taxa de 154,143.

Figura 5: Taxa por 100 mil habitantes do crime roubo e furto de veículos, nas 27 unidades federativas do Brasil no ano de 2021.[5]

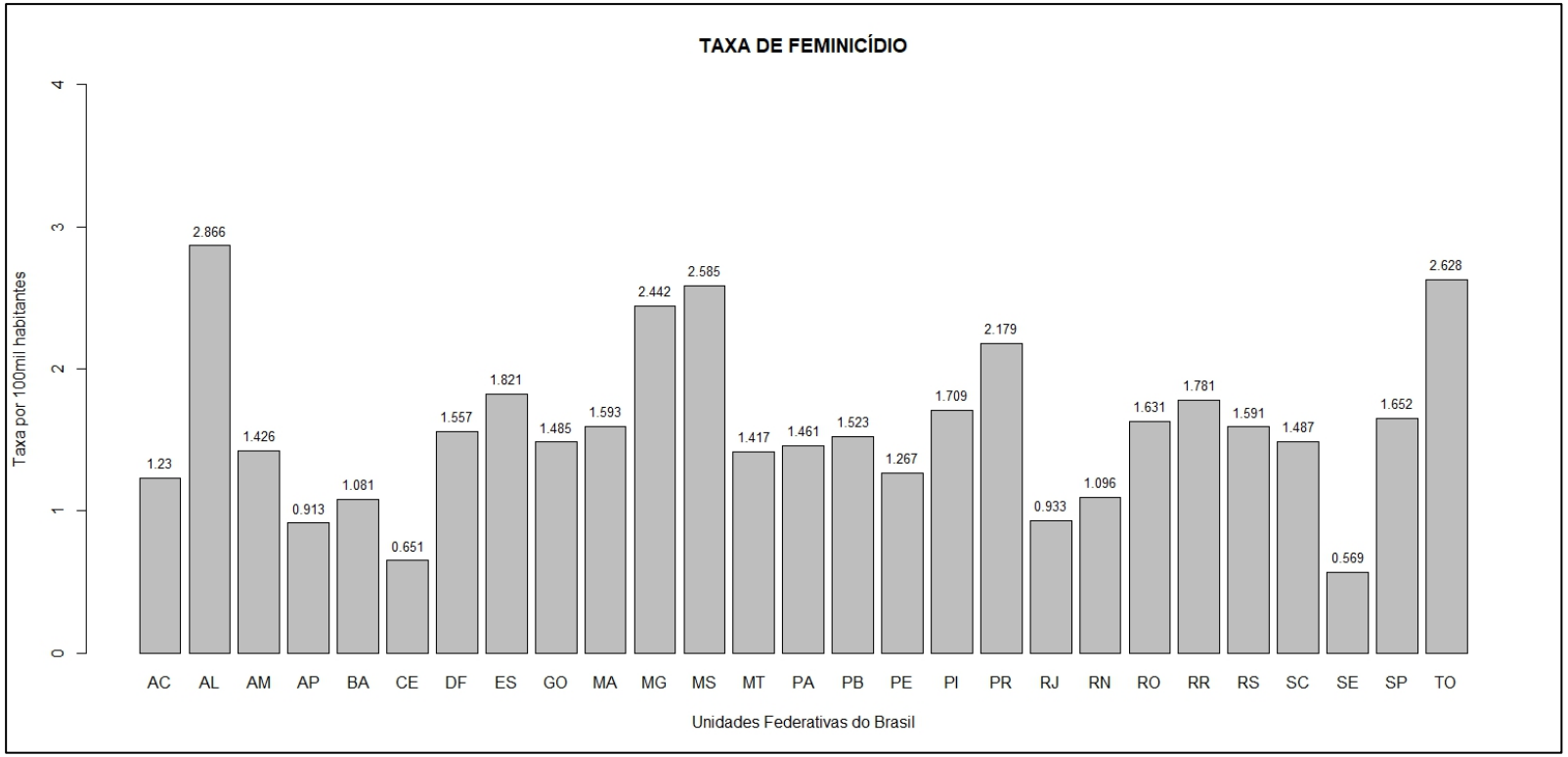


Fonte: A autora.

[5] AC – Acre, AL – Alagoas, AM – Amazonas, AP – Amapá, BA – Bahia, CE – Ceará, DF – Distrito Federal, ES – Espírito Santo, GO – Goiás, MA – Maranhã, MG – Minas Gerais, MS – Mato Grosso do Sul, MT – Mato Grosso, PA – Pará, PB – Paraíba, PE – Pernambuco, PI – Piauí, PR – Paraná, RJ – Rio de Janeiro, RN – Rio Grande do Norte, RO – Rondônia, RR – Roraima, RS – Rio Grande do Sul, SC – Santa Catarina, SE – Sergipe, TO – Tocantins.

O gráfico apresentado na figura 6 é referente ao crime feminicídio (X6), sendo que a maior incidência ocorreu em Alagoas com uma taxa de 2,866 a cada 100 mil habitantes seguido pelo Tocantins com uma taxa de 2,628 e Sergipe é o estado com menor taxa de ocorrência sendo 0,569 a cada 100 mil habitantes.

Figura 6: Taxa por 100 mil habitantes do crime feminicídio, nas 27 unidades federativas do Brasil no ano de 2021.[6]

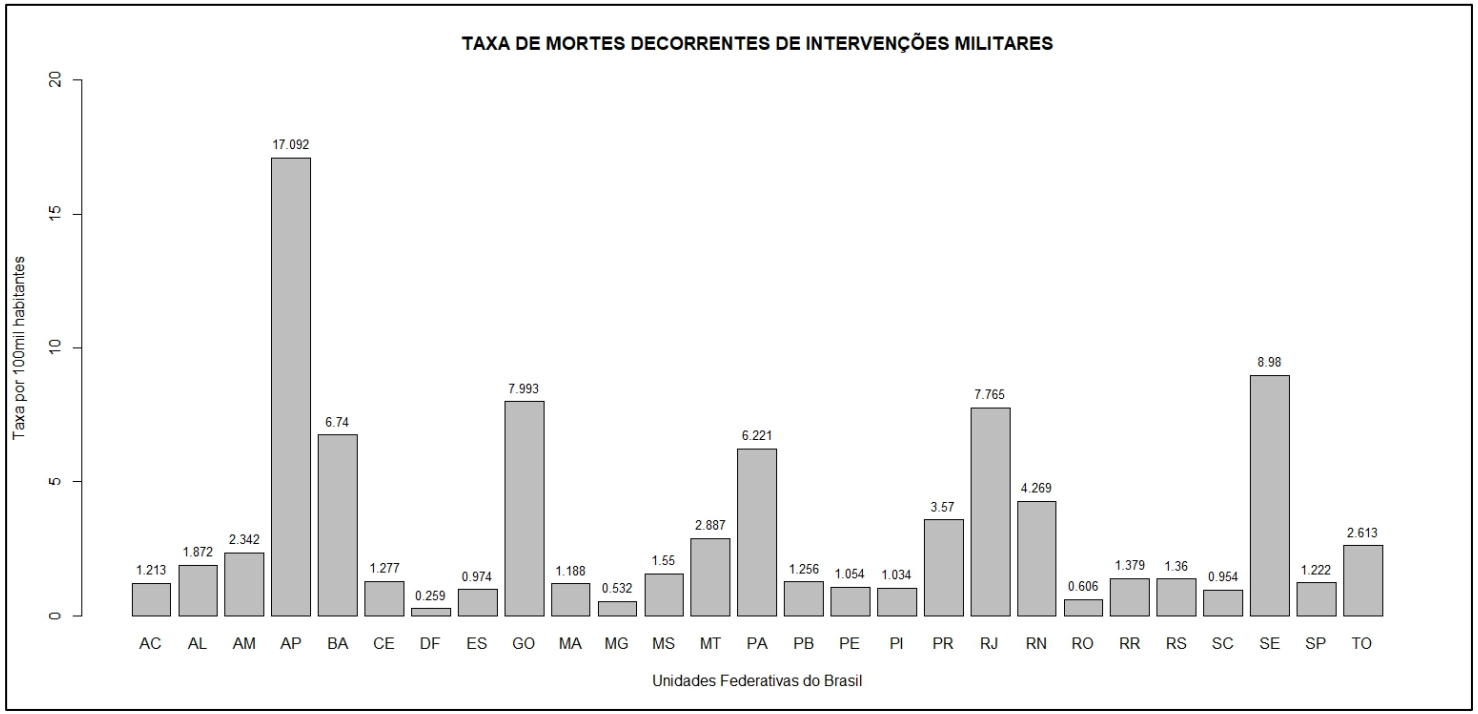


Fonte: A autora.

[6] AC – Acre, AL – Alagoas, AM – Amazonas, AP – Amapá, BA – Bahia, CE – Ceará, DF – Distrito Federal, ES – Espírito Santo, GO – Goiás, MA – Maranhã, MG – Minas Gerais, MS – Mato Grosso do Sul, MT – Mato Grosso, PA – Pará, PB – Paraíba, PE – Pernambuco, PI – Piauí, PR – Paraná, RJ – Rio de Janeiro, RN – Rio Grande do Norte, RO – Rondônia, RR – Roraima, RS – Rio Grande do Sul, SC – Santa Catarina, SE – Sergipe, TO – Tocantins.

Na figura 7 tem-se o gráfico referente ao crime mortes decorrentes de intervenções militares (X7). Observa-se que a maior incidência ocorreu em Amapá com uma taxa de 17,092 a cada 100 mil habitantes seguido pelo Sergipe com 8,980 e o Distrito Federal possui a menor taxa de ocorrência sendo 0,259 a cada 100 mil habitantes

Figura 7: Taxa por 100 mil habitantes do crime mortes decorrentes de intervenções militares, nas 27 unidades federativas do Brasil no ano de 2021.[7]



Fonte: A autora.

[7]AC – Acre, AL – Alagoas, AM – Amazonas, AP – Amapá, BA – Bahia, CE – Ceará, DF – Distrito Federal, ES – Espírito Santo, GO – Goiás, MA – Maranhã, MG – Minas Gerais, MS – Mato Grosso do Sul, MT – Mato Grosso, PA – Pará, PB – Paraíba, PE – Pernambuco, PI – Piauí, PR – Paraná, RJ – Rio de Janeiro, RN – Rio Grande do Norte, RO – Rondônia, RR – Roraima, RS – Rio Grande do Sul, SC – Santa Catarina, SE – Sergipe, TO – Tocantins.

Os gráficos acima evidenciam a disparidade entre os estados brasileiros quanto a criminalidade. Como exemplo, pode-se citar o crime de homicídio lodoso que apresentou a Bahia e o Ceará como os estados com as maiores taxas desse crime, enquanto São Paulo apresentou a menor taxa (Figura 1). Os estados do Amapá e Roraima, apresentaram as maiores taxas de latrocínio e lesão corporal, enquanto Santa Catarina mostrou a menor taxa (Figura 2).

Assim como os demais crimes analisados, as mortes decorrentes de intervenções militares também mostraram essa disparidade, tendo os estados de Amapá e Sergipe com as maiores taxas e o Distrito Federal com a menor taxa (Figura 7). Essa disparidade pode ser atribuída a uma série de fatores, como diferenças socioeconômicas, estruturais e culturais entre os estados.

Ao analisar taxas de criminalidade de diferentes crimes não é incomum encontrar essa diferença entre os estados brasileiros. Segundo o FBSP (2023), em 2022, os estados do Amapá, Bahia e Amazonas aparecem como os estados com as maiores taxas de MVI, enquanto São Paulo, Santa Catarina e o Distrito Federal apresentaram as menos taxas de MVI.

Esses resultados sugerem a necessidade de políticas públicas direcionadas para enfrentar os desafios específicos de cada região, visando a redução da criminalidade e a promoção da segurança pública em todo o país.

O valor mínimo, o valor máximo, a média, o primeiro quartil, a mediana, o terceiro quartil e o desvio padrão para cada variável em estudo é apresentado na tabela 1.

Tabela 1: Estatísticas descritivas dos crimes.[8]

Crimes	Mínimo	Máximo	Média	Primeiro Quartil	Mediana	Terceiro Quartil	Desvio Padrão
X1	6,103	36,916	23,037	17,983	23,247	30,674	8,619
X2	0,245	2,393	0,909	0,550	0,743	1,233	0,524
X3	0,062	1,481	0,315	0,164	0,246	0,344	0,290
X4	13,520	86,470	38,250	23,140	33,740	51,270	19,645
X5	154,100	531	314	200	341,800	387,800	112,610
X6	0,569	2,867	1,577	1,248	1,522	1,745	0,573
X7	0,259	17,091	3,267	1,121	1,379	3,919	3,754

Fonte: A autora.

Observa-se na tabela 1, que o crime roubo e furto de veículos (X5) apresentou média de 314, sendo o maior valor médio quando comparado aos outros crimes apresentados na tabela 1. Dos sete crimes estudados o que apresentou menor média foi lesão corporal seguida de morte (X3).

Foi realizada a análise de agrupamento para as variáveis padronizadas. Utilizou-se diferentes distâncias e métodos de agrupamento e foi calculada a correlação cofenética para cada situação, conforme tabela 2.

[8] X1 – homicídio doloso em número de vítimas; X2 – latrocínio; X3 – lesão corporal seguida de morte; X4 – estupro (incluindo estupro de vulnerável); X5 – roubo e furto de veículos; X6 – feminicídio; X7 – mortes decorrentes de intervenções militares.

Tabela 2: Correlação cofenética para os métodos e distâncias considerados sobre os dados de criminalidade.

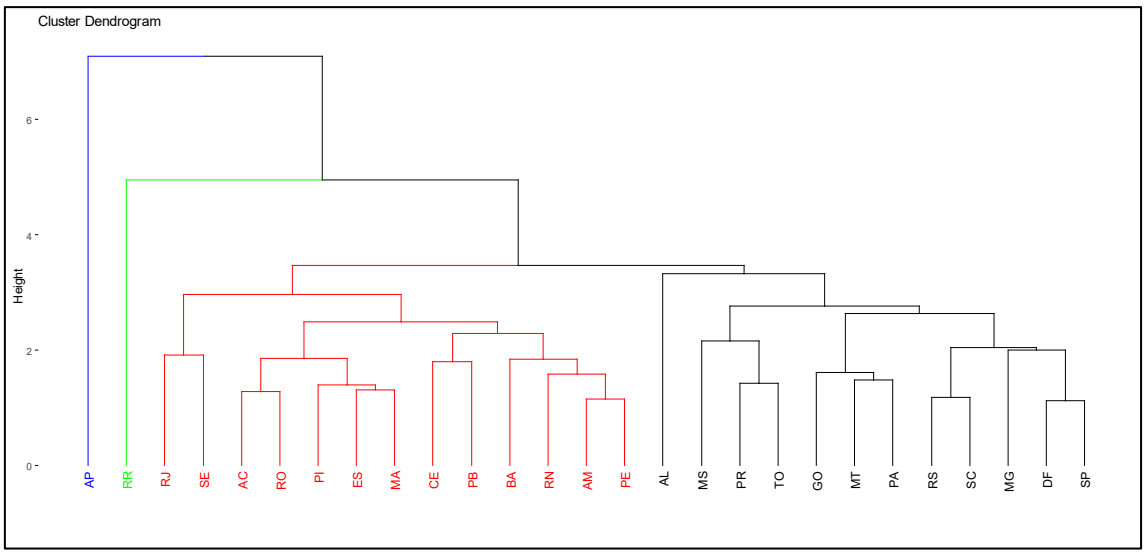
Métodos de agrupamentos	Distâncias	Correlação cofenética
ligação individual	distância euclidiana	0,851
ligação individual	distância de Manhattan	0,828
ligação individual	distância de Mahalanobis	0,711
ligação completa	distância euclidiana	0,807
ligação completa	distância de Manhattan	0,795
ligação completa	distância de Mahalanobis	0,682
ligação média	distância euclidiana	0,880
ligação média	distância de Manhattan	0,867
ligação média	distância de Mahalanobis	0,769
Centróide	distância euclidiana	0,850
Centróide	distância de Manhattan	0,833
Centróide	distância de Mahalanobis	0,760
Ward	distância euclidiana	0,629
Ward	distância de Manhattan	0,687
Ward	distância de Mahalanobis	0,536

Fonte: A autora.

A correlação cofenética foi maior para o método ligação média com a distância euclidiana. Considerando esse método e essa distância, foi utilizado o pacote NbClust para definir o número de cluster a ser utilizado na partição final. Segundo a maioria dos índices o melhor número de clusters é 4.

Na figura 8 é apresentado o dendrograma, obtido pelo método de agrupamento hierárquico da ligação média usando a distância euclidiana. A tabela 3 mostra os quatro grupos formados.

Figura 8: Dendrograma obtido pelo método de agrupamento hierárquico da ligação média usando a distância euclidiana.[9]



Fonte: A autora.

Tabela 3: Grupos formados

Grupos	Estados
1	Amapá
2	Roraima
3	Rio de Janeiro, Sergipe, Acre, Rondônia, Piauí, Espírito Santo, Maranhão, Ceará, Paraíba, Bahia, Rio Grande do Norte, Amazonas, Pernambuco
4	Alagoas, Mato Grosso do Sul, Paraná, Tocantins, Goiás, Mato Grosso e Pará, Rio Grande do Sul, Santa Catarina, Minas Gerais, Distrito Federal e São Paulo

Fonte: A autora.

[9] AC – Acre, AL – Alagoas, AM – Amazonas, AP – Amapá, BA – Bahia, CE – Ceará, DF – Distrito Federal, ES – Espírito Santo, GO – Goiás, MA – Maranhã, MG – Minas Gerais, MS – Mato Grosso do Sul, MT – Mato Grosso, PA – Pará, PB – Paraíba, PE – Pernambuco, PI – Piauí, PR – Paraná, RJ – Rio de Janeiro, RN – Rio Grande do Norte, RO – Rondônia, RR – Roraima, RS – Rio Grande do Sul, SC – Santa Catarina, SE – Sergipe, TO – Tocantins.

Na tabela 4 apresenta-se as médias, para cada grupo, das variáveis em estudo. Observa-se que o grupo 1 apresenta maior média para os crimes X1, X2, X3 e X7. O grupo 2 apresenta maior média para os crimes X4 e X5 e o grupo 4 apresenta maior média para o crime X6.

Tabela 4: Média para cada um dos grupos formados.[10]

Grupos	Média X1	Média X2	Média X3	Média X4	Média X5	Média X6	Média X7
1	32,816	2,393	1,481	60,505	214,533	0,913	17,092
2	31,254	1,992	0,919	83,344	489,079	1,781	1,379
3	27,456	1,008	0,255	29,646	395,460	1,272	2,977
4	16,751	0,587	0,234	41,963	219,482	1,946	2,586

Fonte: A autora.

[10] X1 – homicídio doloso em número de vítimas; X2 – latrocínio; X3 – lesão corporal seguida de morte; X4 – estupro (incluindo estupro de vulnerável); X5 – roubo e furto de veículos; X6 – feminicídio; X7 – mortes decorrentes de intervenções militares.

5. CONCLUSÃO

A análise de agrupamento para os dados de criminalidade, nas 27 unidades federativas do Brasil para o ano de 2021, mostrou-se eficiente, formando 4 grupos com correlação cofenética igual a 0,88. O Amapá e Roraima formaram dois grupos com apenas um estado cada. Tem-se o Rio de Janeiro, Sergipe, Acre, Rondônia, Piauí, Espírito Santo, Maranhão, Ceará, Paraíba, Bahia, Rio Grande do Norte, Amazonas e Pernambuco pertencentes ao mesmo grupo e Rio Grande do Sul, Santa Catarina, Minas Gerais, Distrito Federal e São Paulo formando outro grupo. Para trabalhos futuros pode-se aplicar outras técnicas de agrupamento aos dados de criminalidade.

6. REFERÊNCIAS

CHARRAD, M.; GHAZZALI, N.; BOITEAU, V.; NIKNAFS, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. **Journal of Statistical Software**, [*S.l.*], v. 61, n. 6, p. 1-36, 2014.

FÁVERO, L. P. et al. **Análise de dados:** modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

FERREIRA, D. F. **Estatística multivariada**. Lavras: Editora UFLA, 2008.

DE FIGUEIREDO, S. O.; SINCORÁ, L. A.; LEITE, M. C. de O.; BRANDÃO, M. M. Fatores determinantes do controle da criminalidade em gestão de políticas de segurança pública. **Revista de Administração Pública**, Rio de Janeiro, v. 55, n. 2, p. 438-458, mar./abr. 2021.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA. 16º Anuário Brasileiro de Segurança Pública. *In:* Fórum Brasileiro de Segurança Pública. **Fórum Brasileiro de Segurança Pública.** São Paulo, 2022. Disponível em: https://apidspace.universilab.com.br/server/api/core/bitstreams/c0c2a9ec-d322-487a-b54f-a305cb736798/content. Acesso em: 20 fev. 2024.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA. 17º Anuário Brasileiro de Segurança Pública. *In:* Fórum Brasileiro de Segurança Pública. **Fórum Brasileiro de Segurança Pública.** São Paulo, 2023. Disponível em: https://forumseguranca.org.br/wp-content/uploads/2023/07/anuario-2023.pdf. Acesso em: 20 fev. 2024.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA. [*S. l.*], 2024. Portal. Disponível em: https://publicacoes.forumseguranca.org.br/items/4f923d12-3cb2-4a24-9b63-e41789581d30/full. Acesso em: 20 fev. 2024.

GUIMARÃES, F. T.; BECKER, K. L. A Criminalidade no Rio Grande do Sul: Análise Exploratória de Dados Espaciais para os Anos de 2002, 2010 e 2018. **Revista Economia Ensaios**, Uberlândia, v. 36, n. 2, p. 79-109, jul./ dez. 2021.

KASSAMBARA, A.; MUNDT, F. **factoextra:** Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7, 2020. Disponível em: https://CRAN.R-project.org/package=factoextra. Acesso em: 08 mar. 2024.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2013.

NASCIMENTO, J. M. P. **Análise de agrupamentos aplicada aos dados de violência no Brasil.** 2019. Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Federal de Uberlândia, Uberlândia, 2019.

R CORE TEAM. **R:** A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2024. Disponível em: https://www.R-project.org/. Acesso em: 29 fev. 2024.

REIS, E. **Estatística multivariada**. 2. ed. Lisboa: Sílabo, 2001.

SARAÇLI, S.; AKŞIT, M. Comparison of Hierarchic Clustering Methods with Cophenetic Correlation Coefficient in Big Data. **Afyon Kocatepe University Journal of Science and Engineering**, v. 22, n. 3, p. 552-559, 2022.

DA SILVA, A. R. **biotools:** Tools for Biometry and Applied Statistics in Agricultural Science. R package version 4.2, 2021. Disponível em: https://cran.r-project.org/package=biotools. Acesso em: 08 mar. 2024.

DA SILVA, A. R. **Métodos de análise multivariada em R**. Piracicaba: FEALQ, 2016.

DA SILVA, A. R.; DIAS, C. T. dos S. A cophenetic correlation coefficient for Tocher's method. **Pesquisa Agropecuária Brasileira**, [*S.l.*], v. 48, n. 6, p. 589-596, 2013.

DA SILVA, A. R.; MALAFAIA, G.; MENEZES, I. P. P. biotools: an R function to predict spatial gene diversity via an individual-based approach. **Genetics and Molecular Research,** [*S.l.*], v. 16, n. 2, p. 1- 6, 2017.

VENTORIM, F. C.; NETTO, V. M. Criminalidade e espaço urbano: as redes de relação entre crime, vítimas e localização no Rio de Janeiro. **urbe. Revista Brasileira de Gestão Urbana**, [*S.l.*], v. 15, p. 1-18, 2023.