Universidade Federal de Uberlândia

Instituto de Ciências Exatas e Naturais do Pontal

Curso de Matemática

Trabalho de Conclusão de Curso

Análise de componentes principais: dados de criminalidade no Brasil

por

Maria Victória Paulino de Souza

Licenciatura em Matemática – Ituiutaba - MG

Orientador: Profa. Dra. Franciella Marques da Costa

Análise de componentes principais: dados de criminalidade no Brasil

Este exemplar corresponde à redação final da Monografia devidamente corrigida e defendida por **Maria Victória Paulino de Souza** e aprovada pela comissão julgadora.

Ituiutaba, 17 de dezembro de 2018.

Profa. Dra. Franciella Marques da Costa

Banca Examinadora:

Profa. Dra. Franciella Marques da Costa.
Profa. Dra. Kátia Gomes Facure Giaretta.
Prof. Luiz Fernando Silva Resende.

Monografia apresentada ao Instituto de Ciências Exatas e Naturais do Pontal, UFU como requisito parcial para obtenção do título de Licenciada em Matemática.

Dedico este trabalho aos meus pais, José Onaldo e Maria Aparecida, pois sem eles este trabalho e muitos dos meus sonhos não se realizariam.

AGRADECIMENTOS

Agradeço primeiramente a Deus por ser essencial em minha vida, autor de meu destino, meu guia, socorro presente na hora da angústia.

Aos meus pais Maria Aparecida e José Onaldo, pelo incentivo, por terem acreditado em mim e não medirem esforços para realizar todos os meu sonhos e sempre me mostrarem que não estou sozinha nesta caminhada.

Agradeço meus irmãos, Júnior, Marcos e Márcio e minha irmã Elaine, por cuidarem tão bem de mim, por todos os conselhos e pelo apoio durante toda esta jornada.

Obrigada a todos os meus familiares, da família Rodrigues e família Souza, que estiveram ao meu lado torcendo e me incentivando a seguir sempre em frente.

O meu muito obrigada a família que constitui em Ituiutaba, República UFUracão e agregados, pela amizade sincera, por terem aguentado todos meus momentos de fragilidade, nunca me deixando cair e por terem sido meu porto seguro nesta cidade.

Agradeço aos amigos, amigas e minha amada Atlética XVII de Julho, pelo apoio em cada momento e por terem feito desta, a etapa e os melhores anos da minha vida.

Obrigada a minha orientadora Franciella Marques, por ter confiado em mim e acreditado em meu potencial. Obrigada pela paciência e inúmeros conselhos, por todo o tempo que dedicou a me ajudar durante o processo de realização deste trabalho.

A esta universidade, ao curso de Matemática, aos docentes, diretores, coordenadores e administração que proporcionaram o melhor dos ambientes para que esse trabalho fosse realizado. Agradeço também ao PET Matemática Pontal, por ter auxiliado no meu processo de formação e hoje poder notar o progresso que obtive durante a graduação.

Deixo aqui o meu muito obrigado a todos que direta ou indiretamente participaram desse processo.

RESUMO

Segurança pública é um tema bastante discutido no Brasil visto que o país enfrenta problemas extremamente graves com relação à criminalidade e violência. A sensação de insegurança está presente diariamente na vida de grande parte da população brasileira, principalmente nos grandes centros. O objetivo deste trabalho é analisar dados de criminalidade, dos 26 estados brasileiros e o distrito federal, por meio da análise de componentes principais. Utilizou-se as variáveis homicídio doloso, latrocínio, estupro, tentativa de estupro, roubo de veículos e furto de veículos. Para realizar as análises estatísticas será utilizado o software R conjuntamente com as bibliotecas MVN, psych, FactoMineR e factoextra. Inicialmente foi realizada uma análise estatística descritiva, em seguida testou-se a normalidade dos dados por meio da aplicação do teste de Shapiro-Wilk para normalidade univariada e o teste multivariado Shapiro-Wilk de Royston. Com o objetivo de verificar se os dados são correlacionados utilizou-se o teste de esfericidade de Bartlett e por fim aplicou-se as técnicas de análise de componentes principais. A análise de componentes principais para dados de criminalidade mostrou-se satisfatória, pois reduziu a quantidade de 6 variáveis para duas componentes principais que explicam aproximadamente 71,14% de toda variação dos dados. Também possibilitou a criação de um índice de criminalidade que permitiu classificar o estado quanto aos crimes homicídio doloso, latrocínio, estupro, tentativa de estupro, roubo de veículos e furto de veículos.

Palavras-chave: Componentes principais. Criminalidade. Estatística multivariada.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Segurança pública é um tema bastante discutido no Brasil visto que o país enfrenta problemas extremamente graves com relação à criminalidade e violência. A sensação de insegurança está presente diariamente na vida de grande parte da população brasileira, principalmente nos grandes centros.

Segundo o Anuário Brasileiro de Segurança Pública (2018), o Brasil registrou 63.880 mortes violentas intencionais em 2017, representando uma taxa de 30,8 para cada 100 mil habitantes, sendo que 26,2% do total ocorreram nas capitais. Em 2016 foram registrados 61.283 mortes violentas intencionais sendo 29,7 a taxa por 100 mil habitantes (Anuário Brasileiro de Segurança de Segurança Pública, 2017).

De acordo com o Anuário Brasileiro de Segurança de Segurança Pública (2017), os três estados com maior taxa (por 100 mil habitantes) de mortes violentas intencionais, em 2016, foram Sergipe (64), Rio Grande do Norte (56,9) e Alagoas (55,9).

Foram registrados 60.018 estupros em 2017 e 49.497 em 2016, 82.684 registros de pessoas desaparecidas em 2017 e 71.796 em 2016. Em apenas 10 anos foram pelo menos 694.007 pessoas dadas como desaparecidas, considerando os registros policiais. Foram 1.066.674 veículos roubados ou furtados em 2015 e 2016. Estas informações foram obtidas no Anuário Brasileiro de Segurança Pública (2017) e no Anuário Brasileiro de Segurança Pública (2018).

Métodos estatísticos são utilizados para análise e interpretação de dados em diversas áreas, inclusive para análise de dados de criminalidade. A análise de componentes principais e análise fatorial, métodos da estatística multivariada, foram utilizadas para analisar dados de crimes em 26 estados dos EUA (NEISSE; HONGYU, 2016).

A estatística multivariada visa à análise simultânea de várias variáveis medidas em uma mesma unidade amostral, sendo uma área de estatística de extrema importância utilizada para análise de dados em diversas áreas do conhecimento.

O objetivo deste trabalho é analisar dados de criminalidade, dos 26 estados brasileiros e o distrito federal, por meio da análise de componentes principais. O conjunto de dados utilizado foi obtido no Anuário Brasileiro de Segurança Pública (2017) e são referentes ao ano de 2016.

Este trabalho está organizado em cinco seções: introdução, referencial teórico, material e métodos, resultados e discussão e conclusão. Na introdução apresentam-se informações relevantes em relação à criminalidade no Brasil, no referencial teórico será abordada a teoria referente à análise de componentes principais, no material e métodos descreve-se o conjunto de dados analisado no trabalho, os principais passos para a realização das análises e apresenta o software utilizado. Será descrito e discutido os resultados na seção resultados e discussão e na última seção apresenta-se as conclusões com base nas análises realizadas.

2 REFERENCIAL TEÓRICO

2.1 Teste de esfericidade de Bartlett

Para realizar uma análise de componentes principais é preciso que exista correlação entre variáveis. Quando as variáveis seguem uma distribuição normal multivariada, se a matriz de correlação é uma matriz diagonal implica que as variáveis são independentes (MINGOTI, 2013). O teste de esfericidade de Bartlett é utilizado para testar se a matriz de correlação é uma matriz diagonal. O teste apresentado a seguir está de acordo com o exposto em Lattin, Carroll e Green (2011).

As hipóteses do teste são:

$$H_0: P_{pxp} = I_{pxp}$$
$$H_1: P_{pxp} \neq I_{pxp}$$

em que P_{pxp} é a matriz de correlação populacional e I_{pxp} é a matriz identidade. A estatística de teste é dada por:

$$\chi^2 = -\left[(n-1) - \frac{(2p+5)}{6}\right] ln|R|$$

em que $ln|R|$ é o logaritmo natural do determinante da matriz de correlação, p é o número de variáveis e n é o tamanho amostral. Sob H_0verdadeira a estatística de teste χ^2tem distribuição assintótica qui-quadrado com $(p^2 - p)/2$ graus de liberdade (FERREIRA, 2008).

A matriz de correlação amostral é dada por,

$$R_{pxp} = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1p} \\ R_{21} & R_{22} & \vdots & R_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ R_{p1} & R_{p2} & \cdots & R_{pp} \end{bmatrix}$$

em que o coeficiente de correlação amostral de Pearson entre a i-ésima e a j-ésima variáveis é dado por,

$$R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

em que a variância amostral da i-ésima variável é dada por,

$$S_{ij} = S_{ii} = \frac{\sum_{l=1}^{n}(X_{il} - \bar{X}_i)^2}{n-1}, para\ i = j$$

e a covariância amostral da i-ésima e a j-ésima variáveis é denotada por,

$$S_{ij} = \frac{\sum_{l=1}^{n}(X_{il} - \bar{X}_i)(X_{jl} - \bar{X}_j)}{n-1}, para\ i \neq j.$$

O coeficiente de correlação também pode ser obtido utilizando a fórmula mostrada em Triola (2017),

$$R_{ij} = \frac{n\left(\sum X_i X_j\right) - \left(\sum X_i\right)\left(\sum X_j\right)}{\sqrt{n\left(\sum X_i^2\right) - \left(\sum X_i\right)^2}\sqrt{n\left(\sum X_j^2\right) - \left(\sum X_j\right)^2}}$$

Existem vários testes que podem ser utilizados para testar a normalidade multivariada, dentre eles o teste multivariado Shapiro Wilk de Royston, para maiores informações consultar Ferreira (2008).

2.2 Componentes principais

A teoria apresentada a seguir está de acordo com o exposto em Mingoti (2013).

Considere uma amostra aleatória de tamanho n, em que, para cada elemento da amostra foram observadas p variáveis de interesse. Dessa forma, a matriz de dados é dada por,

$$X_{nxp} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{p1} \\ X_{12} & X_{22} & \vdots & X_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix}$$

em que o primeiro índice é referente à variável e o segundo índice indica a unidade amostral .

A matriz de covariâncias amostrais é definida por,

$$S_{pxp} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \vdots & S_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}$$

em que a variância amostral da i-ésima variável é dada por,

$$S_{ij} = S_{ii} = \frac{\sum_{l=1}^{n}(X_{il} - \bar{X}_i)^2}{n-1}, para\ i = j$$

e a covariância amostral da i-ésima e a j-ésima variáveis é denotada por,

$$S_{ij} = \frac{\sum_{l=1}^{n}(X_{il} - \bar{X}_i)(X_{jl} - \bar{X}_j)}{n-1}, para\ i \neq j.$$

Os autovalores $\hat{\lambda}_i$, com $i = 1, 2, \cdots, p$, da matriz S_{pxp} são obtidos resolvendo a equação característica dada por,

$$\left|S_{pxp} - \hat{\lambda} I_{pxp}\right| = 0.$$

O autovetor $\hat{v}_i$ correspondente ao autovalor $\hat{\lambda}_i$ é um vetor não nulo dado por,

$$S_{pxp}\hat{v}_{px1} = \hat{\lambda}\hat{v}_{px1}$$

em que,

$$\hat{v}_i = \begin{bmatrix} \hat{v}_{i1} \\ \hat{v}_{i2} \\ \vdots \\ \hat{v}_{ip} \end{bmatrix}$$

Para obter os autovetores normalizados, $\hat{e}_1, \hat{e}_2, \cdots, \hat{e}_p$, basta dividir cada componente do vetor $\hat{v}_i$ pelo seu comprimento que é dado por,

$$\|\hat{v}_i\| = \sqrt{\hat{v}_{i1}^2 + \hat{v}_{i2}^2 + \cdots + \hat{v}_{ip}^2}$$

A j-ésima componente principal amostral obtida via matriz de covariância é definida por,

$$\hat{Y}_j = \hat{e}_{j1}X_1 + \hat{e}_{j2}X_2 + \cdots + \hat{e}_{jp}X_p$$

A variância total explicada pela j-ésima componente principal amostral obtida via matriz de covariância é dada pela expressão,

$$\frac{Var[\hat{Y}_j]}{Variância\ total\ estimada\ de\ X} = \frac{\hat{\lambda}_j}{\sum_{i=1}^{p} \hat{\lambda}_i}$$

em que $Var[\hat{Y}_j]$ é a variância estimada do componente principal $\hat{Y}_j$.

A correlação estimada entre a variável aleatória X_i, com $i = 1, 2, \cdots, p,$ e a $j - ésima$ componente principal amostral obtida via matriz de covariância é dada por,

$$r_{\hat{Y}_j, X_i} = \frac{\hat{e}_{ij}\sqrt{\hat{\lambda}_j}}{\sqrt{S_{ii}}}$$

em que S_{ii} é a variância amostral da variável aleatória X_i.

As componentes principais podem ser obtidas utilizando a matriz de covariâncias ou a matriz de correlação. No caso de haver uma discrepância muito grande entre as variâncias das variáveis observadas, as componentes principais obtidas utilizando a matriz de covariâncias terá pouca utilidade prática, visto que são bastante influenciadas pelas variáveis com maior

variância. Os dados podem ser transformados com o objetivo de minimizar esse problema. Uma das transformações usuais é padronizar cada variável pela sua média e desvio padrão. Obter as componentes principais utilizando as variáveis padronizadas e a matriz de covariâncias é equivalente a utilizar as variáveis originais e a matriz de correlação.

A matriz de correlação amostral é dada por,

$$R_{pxp} = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1p} \\ R_{21} & R_{22} & \vdots & R_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ R_{p1} & R_{p2} & \cdots & R_{pp} \end{bmatrix}$$

em que o coeficiente de correlação amostral de Pearson entre a i-ésima e a j-ésima variáveis é dado por,

$$R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

O coeficiente de correlação também pode ser obtido utilizando a fórmula mostrada em Triola (2017),

$$R_{ij} = \frac{n\left(\sum X_i X_j\right) - \left(\sum X_i\right)\left(\sum X_j\right)}{\sqrt{n\left(\sum X_i^2\right) - \left(\sum X_i\right)^2}\sqrt{n\left(\sum X_j^2\right) - \left(\sum X_j\right)^2}}$$

Os autovalores $\hat{\lambda}_i, \text{com } i = 1,2,\cdots,p,$ da matriz R_{pxp} são obtidos resolvendo a equação característica dada por,

$$\left|R_{pxp} - \hat{\lambda} I_{pxp}\right| = 0.$$

O autovetor $\hat{v}_i$ correspondente ao autovalor $\hat{\lambda}_i$ é um vetor não nulo dado por,

$$R_{pxp}\hat{v}_{px1} = \hat{\lambda}\hat{v}_{px1}$$

em que,

$$\hat{v}_i = \begin{bmatrix} \hat{v}_{i1} \\ \hat{v}_{i2} \\ \vdots \\ \hat{v}_{ip} \end{bmatrix}$$

Para obter os autovetores normalizados, $\hat{e}_1, \hat{e}_2, \cdots, \hat{e}_p$, basta dividir cada componente do vetor $\hat{v}_i$ pelo seu comprimento que é dado por,

$$\|\hat{v}_i\| = \sqrt{\hat{v}_{i1}^2 + \hat{v}_{i2}^2 + \cdots + \hat{v}_{ip}^2}$$

A j-ésima componente principal amostral obtida via matriz de correlação é definida por,

$$\hat{Y}_j = \hat{e}_{j1}Z_1 + \hat{e}_{j2}Z_2 + \cdots + \hat{e}_{jp}Z_p$$

em que,

$$Z_i = \frac{(X_i - \mu_i)}{\sigma_i}$$

e $E(X_i) = \mu_i \; e \; Var(X_i) = \sigma_i^2, \; i = 1, 2, \cdots, p.$

A variância total explicada pela j-ésima componente principal amostral obtida via matriz de correlação é dada pela expressão,

$$\frac{Var[\hat{Y}_j]}{Variância\ total\ estimada\ de\ Z} = \frac{\hat{\lambda}_j}{\sum_{i=1}^{p} \hat{\lambda}_i} = \frac{\hat{\lambda}_j}{p}$$

em que $Var[\hat{Y}_j]$ é a variância estimada do componente principal $\hat{Y}_j$.

A correlação estimada entre a variável aleatória Z_i, com $i = 1, 2, \cdots, p,$ e a $j - ésima$ componente principal amostral obtida via matriz de correlação é dada por,

$$r_{\hat{Y}_j, Z_i} = \hat{e}_{ij}\sqrt{\hat{\lambda}_j}.$$

Obtidas as componentes principais é necessário decidir quantas componentes é preciso reter para conseguir obter as informações relevantes contidas nos dados e ao mesmo tempo ter

um modelo reduzido. Segundo Ferreira (2008) e Mingoti (2013) pode-se manter as componentes principais, obtidas a partir da matriz de correlação, que estão relacionadas aos autovalores $\hat{\lambda}_i \geq 1$. Outro critério é reter as K primeiras componentes capazes de explicar pelo menos 70% da variação total (FERREIRA, 2008). Também pode ser utilizado o gráfico em que plotamos a ordem K dos componentes na abscissa e o seu autovalor na ordenada, este gráfico é denominado de scree plot. Deve-se observar no scree plot o ponto a partir do qual os valores dos autovalores tendem a se estabilizar, geralmente é quando os autovalores se aproximam de zero (MINGOTI, 2013). Segundo Mingoti (2013) a interpretação prática do componente principal é um fator importante na escolha do número de componentes a ser retido.

Segundo Mingoti (2013) para utilizar as componentes principais deve-se calcular os escores das componentes, que são os valores numéricos obtidos para cada elemento amostral.

3 MATERIAL E MÉTODOS

O conjunto de dados utilizado neste trabalho foi obtido através da website www.forumseguranca.org.br/publicacoes/11o-anuario-brasileiro-de-seguranca-publica.

Para a análise estatística foram consideradas as taxas, por 100 mil habitantes, das variáveis homicídio doloso (X1), latrocínio (X2), estupro (X3), tentativa de estupro (X4), roubo de veículo (X5) e furto de veículo (X6). Os dados são referentes ao ano de 2016 e alguns estados não foram considerados na análise por apresentarem variáveis em que os dados não estavam disponíveis. Para o estado da Paraíba, as informações sobre roubo de veículo e furto de veículo não estavam disponíveis, para o Acre não estava disponível informações sobre estupro, tentativa de estupro, roubo de veículo e furto de veículo e na Bahia tentativa de estupro era uma informação não disponibilizada.

A metodologia utilizada para realizar a análise dos dados está descrita nos passos seguintes:

1. análise estatística descritiva;
2. aplicação do teste de Shapiro-Wilk para normalidade univariada;
3. aplicação do teste multivariado Shapiro-Wilk de Royston;
4. aplicação do teste de Bartlett;
5. utilização de técnicas de análise de componentes principais.

Para realizar as análises estatísticas foi utilizado o software R (R CORE TEAM, 2018), conjuntamente com as bibliotecas MVN (KORKMAZ; GOKSULUK; ZARARSIZ, 2014), psych (REVELLE, 2018), FactoMineR (LE; JOSSE; HUSSON, 2008) e factoextra (KASSAMBARA; MUNDT, 2017). No anexo consta a rotina com os comandos do software R utilizada na análise dos dados.

4 RESULTADOS E DISCUSSÃO

A análise estatística descritiva das variáveis em estudo é apresentada na Tabela 1, em que é apresentado a média, o desvio padrão, o valor mínimo, o valor máximo, o primeiro quartil, a mediana e o terceiro quartil para cada variável.

Tabela 1: Análise estatística descritiva das variáreis em estudo

Variáveis	Média	Desvio padrão	Mínimo	Máximo	Primeiro Quartil	Segundo Quartil	Terceiro Quartil
Homicídio doloso (X1)	30,42	12,61	8,21	57,64	20,74	29,74	37,39
Latrocínio (X2)	1,57	0,61	0,55	2,78	0,99	1,50	1,94
Estupro (X3)	27,78	14,43	4,73	54,36	18,31	23,56	38,83
Tentativa de estupro (X4)	3,96	2,05	1,50	10,16	2,58	3,13	5,09
Roubo de veículos (X5)	301,74	162,55	60,91	653,93	173,96	295,68	399,51
Furto de veículos (X6)	263,70	95,24	108,60	434,70	197,50	264,20	328,90

A Tabela 2 apresenta as correlações entre as variáveis em estudo, sendo que o variável homicídio doloso (X1) e latrocínio (X2) são as variáveis que apresentaram maior correlação e as variáveis menos correlacionadas são latrocínio (X2) e furto de veículos (X6).

Tabela 2: Correlação entre as variáveis

	X1	X2	X3	X4	X5	X6
X1	1,00	0,58	-0,30	-0,53	0,55	-0,61
X2	-	1,00	0,06	-0,18	0,35	-0,01
X3	-	-	1,00	0,53	-0,57	0,43
X4	-	-	-	1,00	-0,56	0,32
X5	-	-	-	-	1,00	-0,31
X6	-	-	-	-	-	1,00

Observa-se na Figura 1 o gráfico de dispersão e o coeficiente de correlação entre as variáveis em análise.

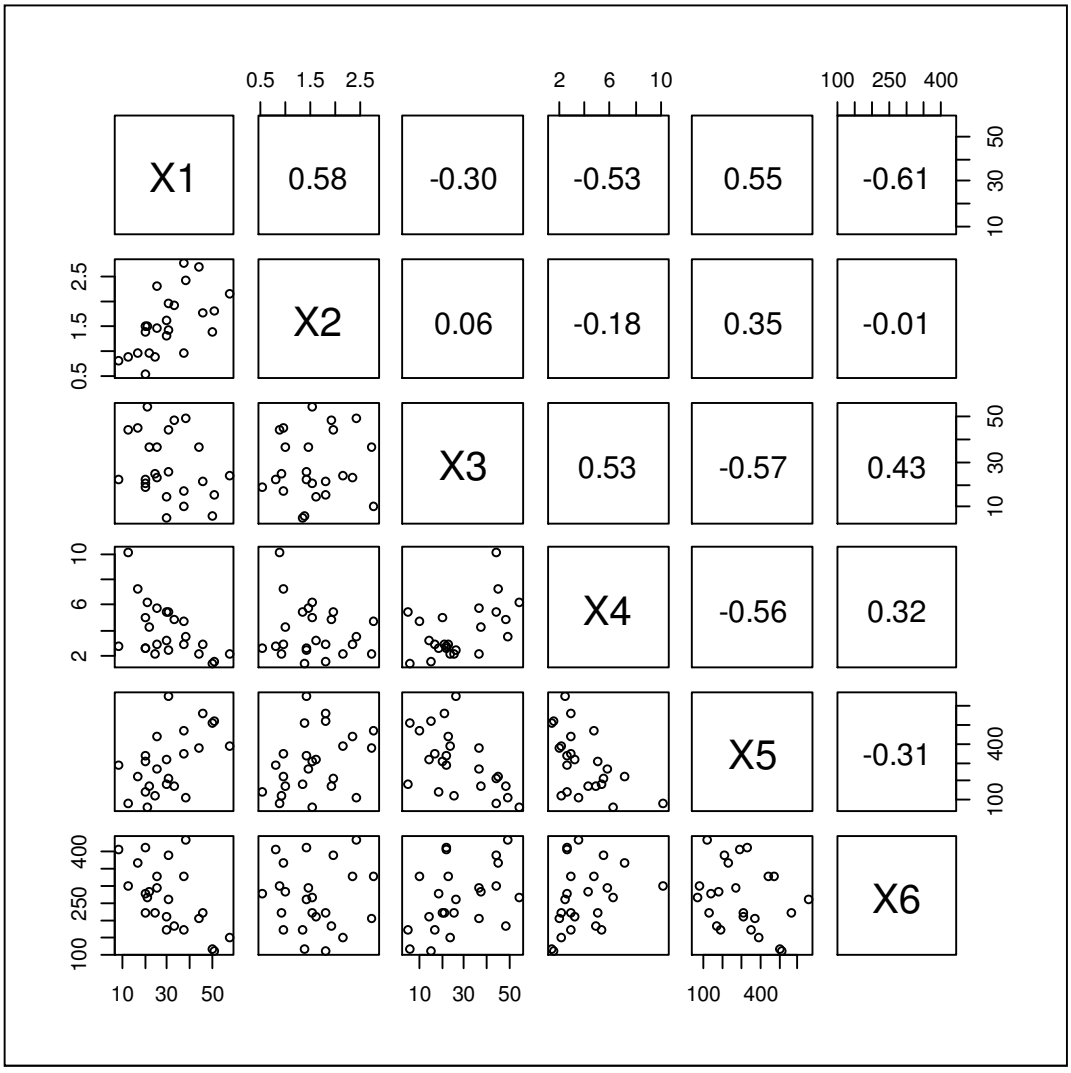

Figura 1: O gráfico de dispersão e o coeficiente de correlação entre as variáveis em análise.

As Figuras 2 e 3 são o boxplot das variáveis originais e o boxplot das variáveis padronizadas pela sua média e desvio padrão, respectivamente. Pode-se observar na Figura 2 que a dispersão dos dados é diferente para todas as variáveis, sendo que a variável roubo de veículos (X5) apresenta maior dispersão e a variável latrocínio (X2) possui menor dispersão. Pela análise da Figura 2 também é possível observar que as variáveis possuem medianas diferentes. Nota-se que a variável tentativa de estupro (X4) possui um outlier.

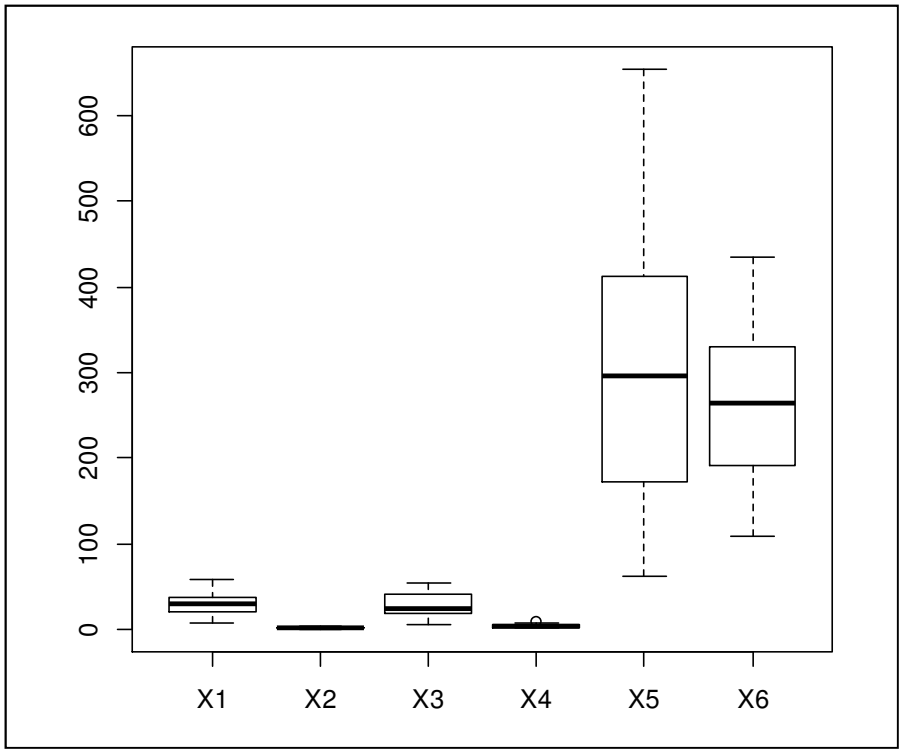

Figura 2: Boxplot das variáveis originais.

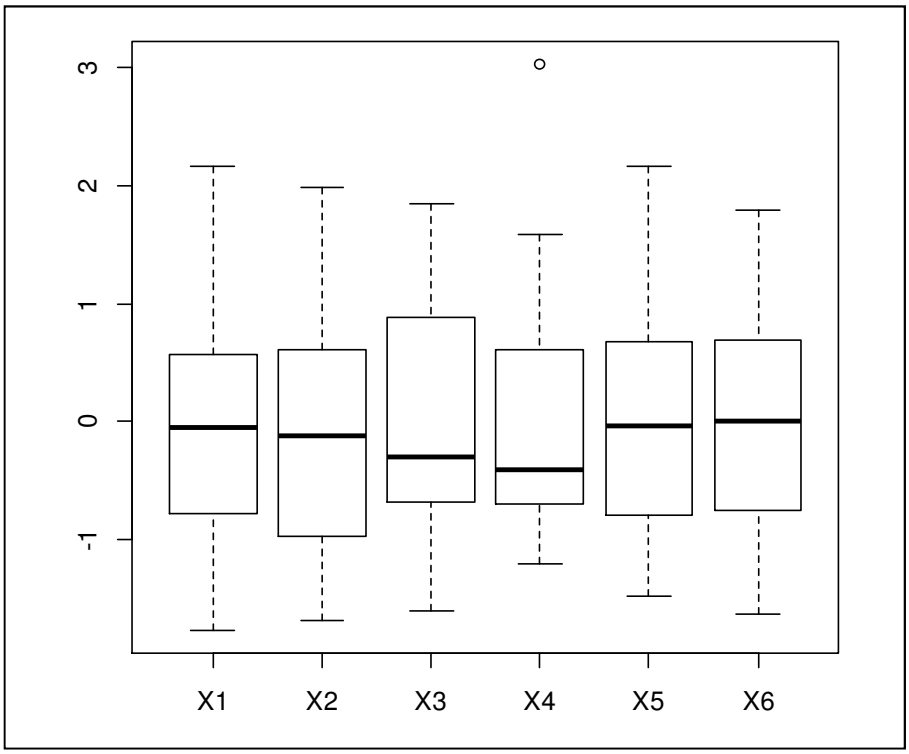

Figura 3: Boxplot das variáveis padronizadas pela sua média e desvio padrão.

A análise do histograma e do Q-Q plot, Figuras 4 e 5, sugerem que as variáveis seguem uma distribuição aproximadamente normal.

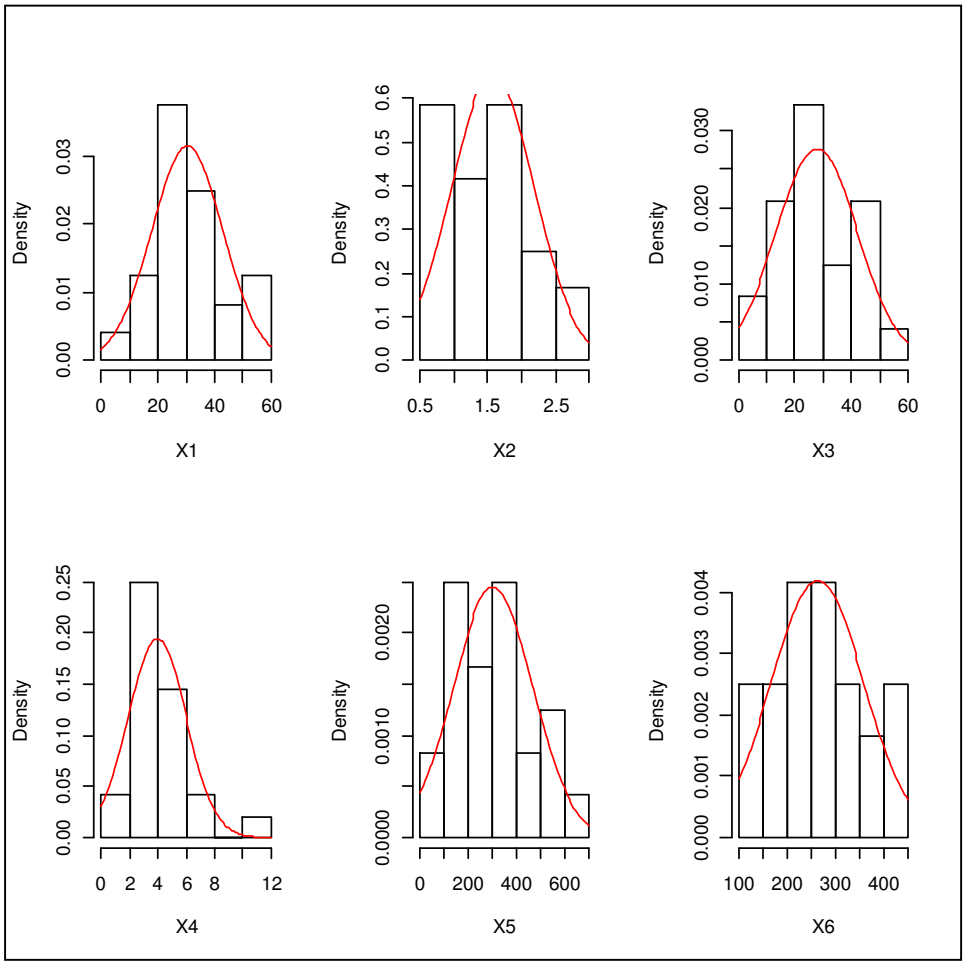

Figura 4: Histograma.

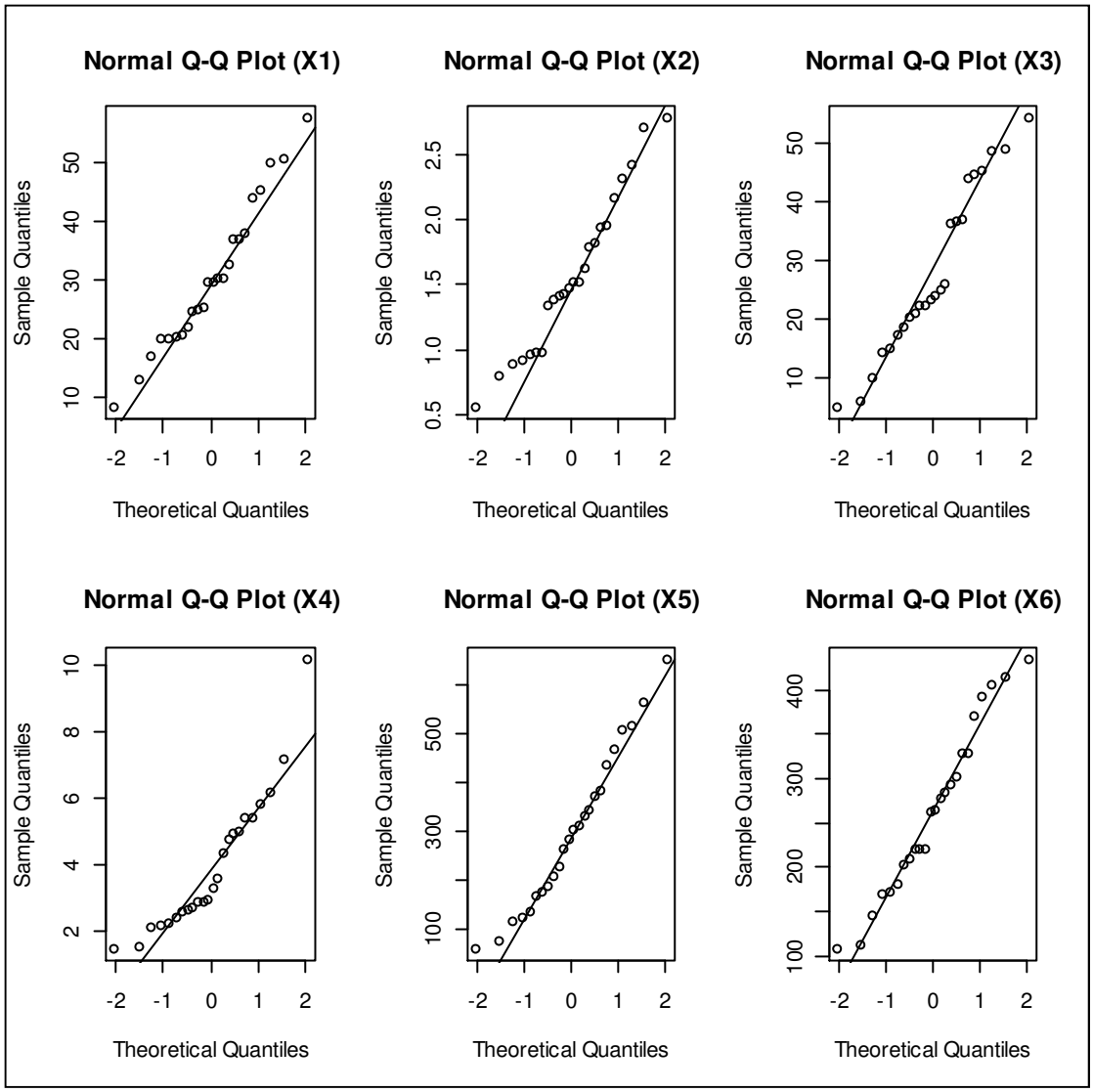

Figura 5: Q-Q plot univariado.

As análises realizadas até o momento indicam que as variáveis seguem uma distribuição aproximadamente normal. A análise do Q-Q plot multivariado, Figura 6, sugere que os dados seguem uma distribuição normal multivariada. Nota-se na Figura 6 a presença de um possível outlier.

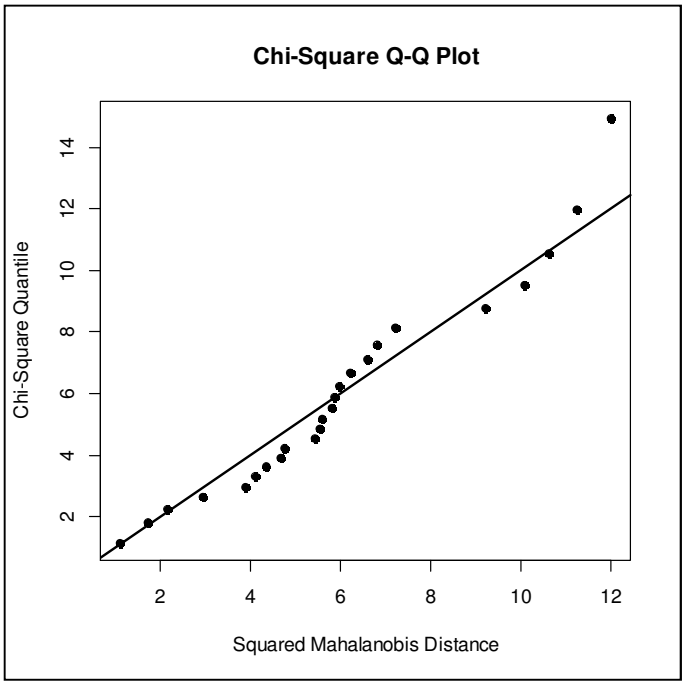

Figura 6: Q-Q plot multivariado.

Como a análise gráfica é bastante subjetiva foi aplicado o teste de Shapiro-Wilk para normalidade univariada e o teste multivariado Shapiro-Wilk de Royston com o objetivo de confirmar os resultados. Os resultados do teste univariado de Shapiro-Wilk são apresentados na Tabela 3. Considerando um nível de significância de 1%, todas as variáveis apresentam distribuição normal, conforme indicou a análise gráfica do histograma e do q-q plot.

Tabela 3: Resultado do teste univariado de Shapiro-Wilk

Variáveis	Estatística do teste	Valor-p
Homicídio doloso (X1)	0,9678	0,6139
Latrocínio (X2)	0,9617	0,4738
Estupro (X3)	0,9424	0,1842
Tentativa de estupro (X4)	0.8851	0,0105
Roubo de veículos (X5)	0,9651	0,5498
Furto de veículos (X6)	0,9634	0,5108

Com a aplicação do teste multivariado Shapiro-Wilk de Royston, obteve-se estatística de teste igual a 8,9354 e $valor - p = 0{,}1228$, concluindo-se que as variáveis seguem uma distribuição normal multivariada, ao nível de significância de 1%.

Com o objetivo de verificar se as variáveis em estudo estão correlacionadas de alguma forma aplicou-se o teste de esfericidade de Bartlett. Com a aplicação do teste obteve-se $\chi^2 = 57{,}4359$ e $valor - p = 6{,}9205\text{x}10^{-07}$, portanto as variáveis não são mutuamente independentes.

Verificado que as variáveis são correlacionadas realizou-se uma análise de componentes principais. Observa-se na Tabela 4 que as componentes 1 e 2 explicam aproximadamente 71,14% de toda variação dos dados, sendo aproximadamente 50,28% do componente principal 1 (CP1) e 20,86 do CP2.

Tabela 4: Componentes principais (CP), autovalores, porcentagem da variância total explicada pelo componente e a porcentagem acumulada da explicação da variância total

Componentes principais	Autovalores	Porcentagem	Porcentagem acumulada
CP1	3,02	50,28	50,28
CP2	1,25	20,86	71,15
CP3	0,83	13,87	85,01
CP4	0,47	7,78	92,79
CP5	0,30	5,03	97,82
CP6	0,13	2,18	100

O scree plot, Figura 7, apresenta a ordem dos k componentes principais e o seu respectivo autovalor. Na Figura 8 observa-se as porcentagens da variância total explicada por cada componente principal.

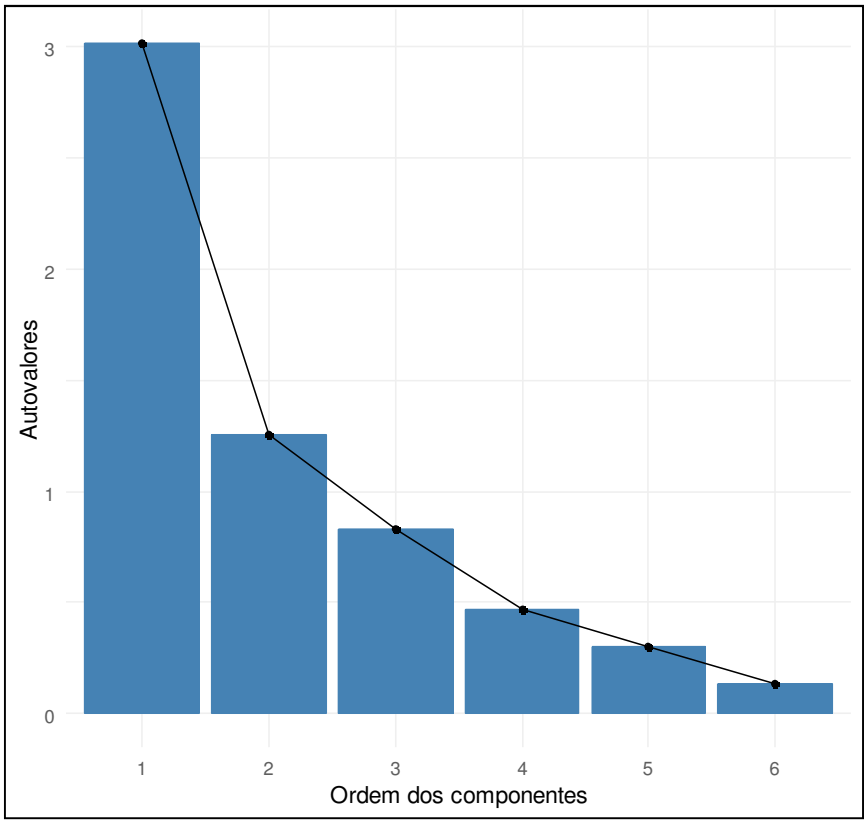

Figura 7: Scree-plot.

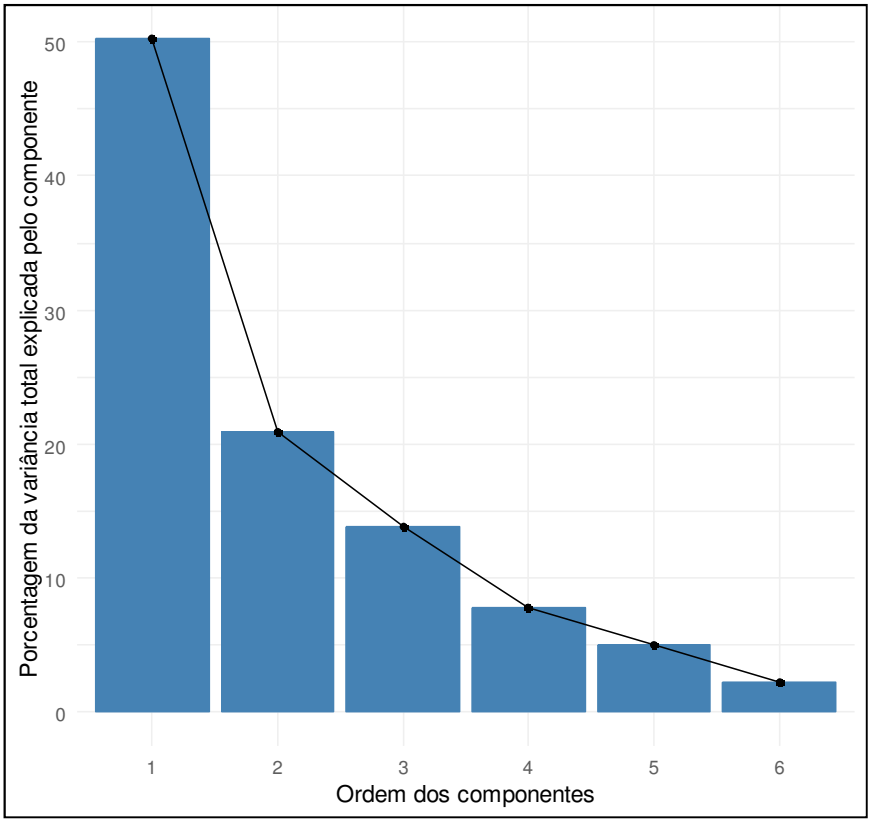

Figura 8: Porcentagem da variância total explicada pelo componente.

Para definir quantos componentes serão retidos foram utilizados os critérios discutidos no referencial teórico. Pode se afirmar com base na avaliação dos autovalores e da porcentagem acumulada da explicação da variância total que CP1 e CP2 resumem de maneira adequada a variação dos dados, visto que explicam pelo menos 70% da variação total e estão associados aos autovalores maiores que 1. Dessa forma, apresentam-se somente as equações para CP1 e CP2,

$$\mathrm{CP1} = -0{,}485Z1 - 0{,}250Z2 + 0{,}392Z3 + 0{,}441Z4 - 0{,}465Z5 + 0{,}371Z6$$

$$\mathrm{CP2} = 0{,}313Z1 + 0{,}758Z2 + 0{,}499Z3 + 0{,}136Z4 + 0{,}010Z5 + 0{,}244Z6$$

A componente principal 1 representa uma comparação das variáveis homicídio doloso (X1), latrocínio (X2), roubo de veículos (X5) com as variáveis estupro (X3), tentativa de estupro (X4), furto de veículos (X6). A CP2 pode ser vista com um índice de criminalidade, que pode ser utilizada para comparar os estados por meio de seus escores calculados para essa componente.

A Tabela 5 e a Figura 9 mostram as correlações entre os dois primeiros componentes e as variáveis. Pode-se dizer que todas as variáveis estão correlacionadas com a componente principal 1 e que existe uma correlação negativa entre as variáveis X1, X2 e X5 e CP1. Roubo de veículo é a variável menos importante no CP2, visto que a correlação entre ela e a componente é próxima de zero.

Tabela 5: Correlação entre CP1, CP2 e as variáveis

Variáveis	CP1	CP2
Homicídio doloso (X1)	-0,84	0,35
Latrocínio (X2)	-0,43	0,85
Estupro (X3)	0,68	0,56
Tentativa de estupro (X4)	0,77	0,15
Roubo de veículos (X5)	-0,81	0,01
Furto de veículos (X6)	0,65	0,27

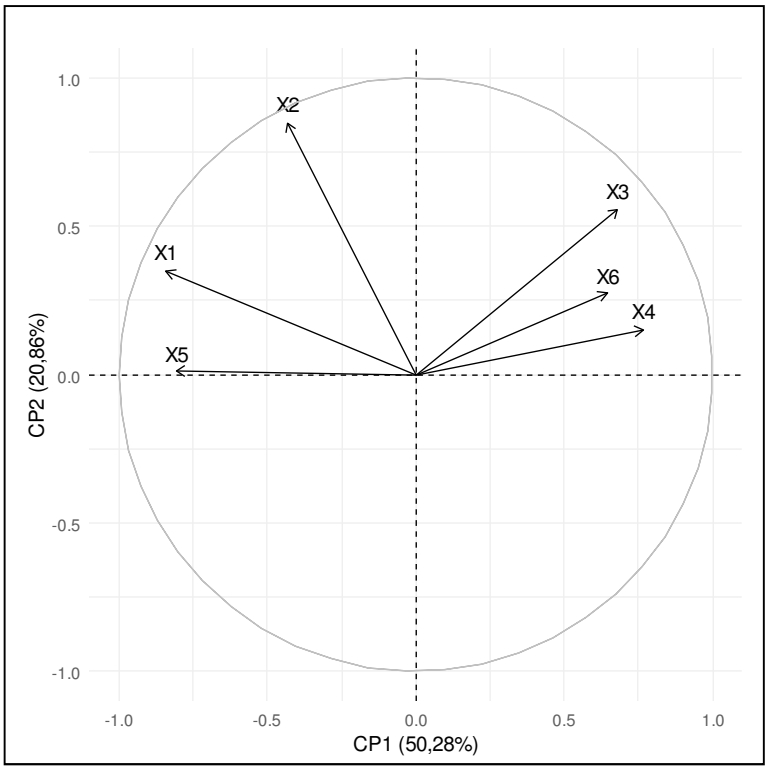

Figura 9: Representação gráfica da correlação entre CP1, CP2 e as variáveis.

Observam-se na Tabela 6 os escores das duas primeiras componentes principais e a classificação dos estados quanto à criminalidade. Como a CP2 pode ser vista com um índice de criminalidade utilizaram-se os escores da CP2 para classificar os estados. Minas Gerais é o estado com menor índice de criminalidade e Amapá é o estado com maior índice. Os escores dos estados também podem ser visualizados na representação gráfica Biplot, Figura 10, na qual se pode verificar que os Estados que apresentam índices altos de criminalidade estão acima do zero com relação ao CP2, como é o caso do 21 que corresponde ao Estado do Amapá. Os estados com baixos índices de criminalidade têm seus pontos abaixo do zero, como é caso do estado de Minas Gerais e São Paulo por exemplo (números 6 e 19).

Tabela 6: Escores das duas primeiras componentes principais e a classificação dos estados, utilizando os escores da CP2, quanto à criminalidade

Codificação dos Estados	Estados	Escores da CP1	Escores da CP2	Posição (CP2)
1	Alagoas	-3,03	-0,19	13
2	Amazonas	-0,61	0,77	17
3	Ceará	-1,04	-1,27	5
4	Espírito Santo	-0,22	-1,29	4
5	Mato Grosso	0,61	1,11	20
6	Minas Gerais	0,83	-1,95	1
7	Pará	-1,63	1,82	23
8	Paraná	1,37	-0,56	9
9	Pernambuco	-2,04	0,26	15
10	Piauí	0,26	-0,63	7
11	Rio de Janeiro	-1,35	-0,32	11
12	Rio Grande do Norte	-3,05	-1,06	6
13	Santa Catarina	3,61	-0,20	12
14	Distrito Federal	0,54	-0,35	10
15	Goiás	-1,31	1,32	21
16	Maranhão	-0,76	-0,61	8
17	Mato Grosso do Sul	2,33	0,78	18
18	Rio Grande do Sul	1,12	0,27	16
19	São Paulo	1,39	-1,44	2
20	Sergipe	-2,52	0,89	19
21	Amapá	1,08	2,45	24
22	Rondônia	1,39	1,50	22
23	Roraima	2,63	0,01	14
24	Tocantins	0,38	-1,32	3

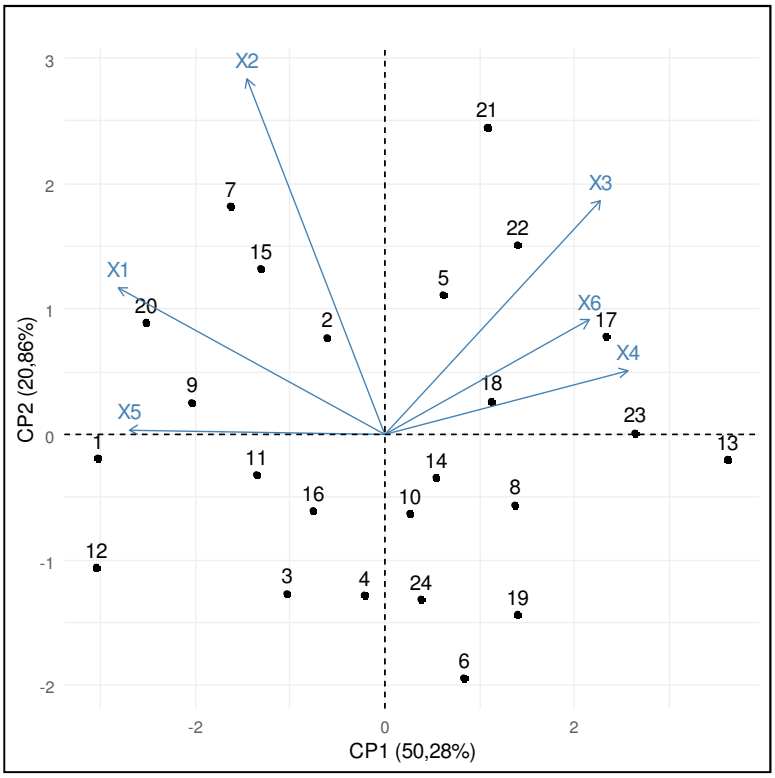

Figura 10: Biplot.

5 CONCLUSÃO

A análise de componentes principais para dados de criminalidade mostrou-se satisfatória, pois reduziu a quantidade de 6 variáveis para duas componentes principais que explicam aproximadamente 71,14% de toda variação dos dados. Também possibilitou a criação de um índice de criminalidade que permitiu classificar o estado quanto aos crimes homicídio doloso, latrocínio, estupro, tentativa de estupro, roubo de veículos e furto de veículos.

REFERÊNCIAS BIBLIOGRAFICAS

BRASIL. Secretaria de Segurança Pública. **Anuário Brasileiro de Segurança Pública 2018.** Disponível em: < http://www.forumseguranca.org.br/atividades/anuario/> Acessado em: 01 jul. 2018

BRASIL. Secretaria de Segurança Pública. **11º Anuário Brasileiro de Segurança Pública.** Disponível em: http://www.forumseguranca.org.br/publicacoes/11o-anuario-brasileiro-de-seguranca-publica/ Acessado em: 01 jul. 2018

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

KASSAMBARA, A.; MUNDT, F. (2017). **Factoextra: Extract and Visualize the Results of Multivariate Data Analyses.** R package version 1.0.5. Disponível em: < https://CRAN.R-project.org/package=factoextra> .

KORKMAZ, S.; GOKSULUK, D.; ZARARSIZ, G. **MVN: An R Package for Assessing Multivariate Normality.** The R Journal. 2014 6(2): 151-162.

LATTIN, J. M.; CARROLL, J. D.; GREEN, P. E. **Análise de dados multivariados.** São Paulo: Cengage Learning, 2011.

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

NEISSE, A. C.; HONGYU, K. Aplicação de Componentes Principais e Análise Fatorial a Dados Criminais de 26 Estados dos Eua. **E&S - Engineering and Science**, v. 2, 2016.

REVELLE, W. (2018) **Psych: Procedures for Personality and Psychological Research**. Northwestern University, Evanston, Illinois, USA. Disponível em: <https://CRAN.R-project.org/package=psych Version = 1.8.10>.

R Core Team (2018). **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. Disponível em: < https: // www.R-project.org/>.

LE, S.; JOSSE J.; HUSSON, F. (2008). **FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software.** 25(1), 1-18. 10.18637/jss.v025.i01.

ANEXO

Rotina com os comandos do software R.

```
install.packages("MVN",dependencies=TRUE)
install.packages("psych",dependencies=TRUE)
install.packages("FactoMineR",dependencies=TRUE)
install.packages("factoextra",dependencies=TRUE)

citation()
citation("MVN")
citation("psych")
citation("FactoMineR")
citation("factoextra")

dados_criminalidade <-read.table("dados.txt",header=T)
dados_criminalidade
dados <- dados_criminalidade[, -1]
dados

summary(dados)
sd(dados$X1)
sd(dados$X2)
sd(dados$X3)
sd(dados$X4)
sd(dados$X5)
sd(dados$X6)
```

```
matriz_correlação <- round(cor(dados), 2)
matriz_correlação

#função disponível em help(pairs), porém com algumas modificações

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
 usr <- par("usr"); on.exit(par(usr))
 par(usr = c(0, 1, 0, 1))
 r <- round(cor(x, y),2)
 txt <- format(c(r, 0.123456789), digits = digits)[1]
 txt <- paste0(prefix, txt)
 if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
 text(0.5, 0.5, txt, cex = 1.5)
}

pairs(dados,  upper.panel = panel.cor)

boxplot(dados)
boxplot(scale(dados))

# Verificando normalidade

library(MVN)

histograma <- mvn(dados,univariatePlot = "histogram")
qqplot <- mvn(dados,univariatePlot = "qqplot")
qqplot <- mvn(dados, multivariatePlot = "qq")

teste_normalidade <- mvn(data = dados, mvnTest = "royston")
teste_normalidade

#teste de esfericidade de Bartlett

library(psych)

r <- cor(dados)
cortest.bartlett(r,n=24)
```

```
# Análise de componentes principais

library(FactoMineR)

acp.cor <- PCA(dados, scale.unit = T, graph = FALSE)

round(acp.cor$eig,2) #autovalores
round(acp.cor$svd$V,3) #autovetores

library(factoextra)

fviz_eig(acp.cor, choice = "eigenvalue", axes = 1,title="",xlab="Ordem dos componentes", ylab="Autovalores")+ theme_minimal()
fviz_screeplot(acp.cor, ncp=7, title="", xlab="Ordem dos componentes", ylab="Porcentagens da variância total explicada pelo componente")+ theme_minimal()
round(acp.cor$var$cor,2) #correlação entre a variável e o componente principal
fviz_pca_var(acp.cor, title="", xlab="CP1 (50,28%)", ylab="CP2 (20,86%)") + theme_minimal()
fviz_pca_biplot(acp.cor,title="", xlab="CP1 (50,28%)", ylab="CP2 (20,86%)") + theme_minimal()
round(acp.cor$ind$coord,2)#escore
```