E&S Engineering and Science
Scientific Journal of FAET and ICET UFMT

Análise de Componentes Principais: resumo teórico, aplicação e interpretação

Principal Component Analysis: theory, interpretations and applications

[1]Kuang Hongyu, [2]Vera Lúcia Martins Sandanielo, [3]Gilmar Jorge de Oliveira Junior

[1]Doutor em Estatística – USP, Professor do Departamento de Estatística – UFMT(prof.kuang@gmail.com)
[2]Doutor em Estatística – UNESP, Professor do Departamento de Estatística – UFMT(veluma@ufmt.br)
[3]Mestre em Estatística – UFRJ, Professor do Departamento de Estatística – UFMT(gilmarjr@ufmt.br)

Resumo: A análise multivariada de uma forma bem geral refere-se aos métodos estatísticos que analisam simultaneamente múltiplas medidas em cada indivíduo ou objeto sob investigação. Nesse contexto, entre as técnicas de multivariadas, a análise de componentes principais (ACP) é uma das técnicas estatísticas mais utilizadas na análise de dados em diversas áreas do conhecimento, como agronomia, zootecnia, ecologia, florestal, medicina, etc. Assim, discute-se neste artigo a aplicabilidade e a interpretação da análise de componentes principais e com utilização do gráfico *biplot*. Também, faz-se uma discussão quanto à escolha da metodologia mais adequada, levando em consideração a informação requerida e os objetivos do pesquisador.

Palavras-chave: Análise multivariada; *biplot*.

Abstract: The multivariate analysis of a well generally refers to all statistical methods to analyze simultaneously multiple measurements on each individual or object under investigation. In this context, among the multivariate techniques, principal component analysis (PCA) is one of the statistical techniques most commonly used in data analysis in various areas of knowledge, such as agronomy, animal science, ecology, forestry, medicine, etc. Thus, this review article discusses the applicability and interpretation of principal component analysis and use of the biplot graph.Also, discuss the choice of the most appropriate methodology, taking into account the information required and the goals of the investigator.

Keywords: multivariate analysis; biplot.

INTRODUÇÃO

A análise de componentes principais (ACP) é uma técnica multivariada de modelagem da estrutura de covariância. A técnica foi inicialmente descrita por Pearson (1901) e uma descrição de métodos computacionais práticos veio muito mais tarde com Hotelling (1933, 1936) que usou com o propósito determinado de analisar as estruturas de correlação.A ACP é uma técnica estatística de análise multivariada que transforma linearmente um conjunto original de variáveis, inicialmente correlacionadas entre si, num conjunto substancialmente menor de variáveis não correlacionadas que contém a maior parte da informação do conjunto original.

A ACP é a técnica mais conhecida e está associada à ideia de redução de massa de dados, com menor perda possível da informação, contudo é importante ter uma visão conjunta de todas ou quase todas as técnicas da estatística multivariada para resolver a maioria dos problemas práticos, também é associada à ideia de redução de massa de dados, com menor perda possível da informação. Procura-se redistribuir a variação observada nos eixos originais de forma a se obter um conjunto de eixos ortogonais não correlacionados (MANLY, 1986; HONGYU, 2015).

A ACP é uma técnica da estatística multivariada que consiste em transformar um conjunto de variáveis originais em outro conjunto de variáveis de mesma dimensão denominadas de componentes principais. Os componentes principais apresentam propriedades importantes: cada componente principal é uma combinação

linear de todas as variáveis originais, são independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados (JOHNSON; WICHERN, 1998; HONGYU, 2015).

O objetivo principal da análise de componentes principais é o de explicar a estrutura da variância e covariância de um vetor aleatório, composto de p-variáveis aleatórias, por meio de combinações lineares das variáveis originais. Essas combinações lineares são chamadas de componentes principais e são não correlacionadas entre si (SANDANIELO, 2008).

Esta técnica pode ser utilizada para geração de índices e agrupamento de indivíduos. A análise agrupa os indivíduos de acordo com sua variação, isto é, os indivíduos são agrupados segundo suas variâncias, ou seja, segundo seu comportamento dentro da população, representado pela variação do conjunto de características que define o indivíduo, ou seja, a técnica agrupa os indivíduos de uma população segundo a variação de suas características.

De acordo com Kubrusly (2001), para se estabelecer um índice que possibilite ordenar um conjunto de *n* objetos, segundo critério definido por um conjunto de *m* variáveis adequadas, é necessário escolher os pesos ou ponderações das variáveis de tal forma que traduzam a informação contida na variável. Para a construção de um índice como combinação linear de variáveis, é desejável que este contenha o máximo de informação fornecida pelo conjunto de variáveis selecionadas. Um método que cria combinações lineares com máxima variância é a análise de componentes principais (SANDANIELO, 2008).

Segundo Regazzi (2000), apesar das técnicas de análise multivariada terem sido desenvolvidas para resolver problemas específicos, as mesmas podem ser utilizadas para resolver outros tipos de problemas como redução da dimensionalidade das variáveis, agrupar os indivíduos (observações) pelas similaridades, em diversas áreas do conhecimento, por exemplo, agronomia, fitotecnia, zootecnia, ecologia, biologia, psicologia, medicina, engenharia florestal, etc.

Assim, o objetivo deste artigo é fornecer informações sobre a ACP que auxilie pesquisadores na escolha e utilização da técnica para analisar e interpretar resultados.

MATERIAL E MÉTODOS

O conjunto de dados utilizado neste (artigo) foi retirado do SAS (2008) sobre um censo que forneceu informações sobre a taxa de criminalidade de algumas cidades dos Estados Unidos: *New York, Los Angeles, Detroit, Washington, Hartford, Honolulu, Boston, Tueson,Portland, Denver, Chicago, Atlanta, Houston, Dallas, New Orleans e Kansas City* . Todas as análises deste artigo foram feitas por meio de rotinas computacionais implementadas no software R 3.0.1 (R Development Core Team, 2014).

Um conjunto de dados sobre a taxa de criminalidade de algumas cidades dos Estados Unidos, citadas acima foi utilizado para exemplificar a técnica e sete variáveis foram estudadas: X_1: Assassinato; X_2: Estupro; X_3: Roubo; X_4: Assalto; X_5: Arrombamento; X_6: Pequenos furtos e X_7: Roubo de veículos.

A obtenção dos componentes principais é realizada por meio da diagonalização de matrizes simétricas positivas semi-definidas. Então, podem-se calcular os componentes principais facilmente e utilizá-los em diferentes aplicações nas mais variadas áreas científicas. Esta facilidade é função da existência de inúmeros programas capazes de realizar cálculos matriciais para diagonalizar uma matriz simétrica positiva semi-definida. Muitos pesquisadores têm utilizado a análise de componentes principais para resolver problemas como da

multicolinearidade em regressão linear, para estimar fatores, que representam outra técnica multivariada de modelagem da matriz de covariâncias, para realizar a modelagem da interação entre fatores em experimentos sem repetição, estudos de divergência e agrupamento entre genótipos em estudo de genética e melhoramento de plantas e animais, entre outras possibilidades (HONGYU, 2012; JOHNSON; WICHERN, 1998).

A ACP tem como principais vantagens: retirar a multicolinearidade das variáveis, pois permite transformar um conjunto de variáveis originais intercorrelacionadas em um novo conjunto de variáveis não correlacionadas (componentes principais). Além disso, reduz muitas variáveis a eixos que representam algumas variáveis, sendo estes eixos perpendiculares (ortogonais) explicando a variação dos dados de forma decrescente e independente (HONGYU, 2015; REGAZZI, 2000).

As desvantagens são: a sensibilidade a outliers, não recomendada quando se tem duplas ausências (muitos zeros na matriz) e dados ausentes. A ACP também não é recomendada quando se tem mais variáveis do que unidades amostrais. Ao reduzir o número de variáveis, há perda da informação de variabilidade das variáveis originais. Mas que a parte explicada seja o padrão de resposta e a outra parte o ruído, ou seja, erro de medida e redundância. A ACP nem sempre funciona (às vezes mesmo com a redução ainda continua grande). É o caso de variáveis originais pouco correlacionadas, com o caso extremo da **R** = **I**, os componentes principais são as próprias variáveis originais (HONGYU, 2015; REGAZZI, 2000).

Sejam as variáveis $X_1, X_2, \ldots, X_p$em cada um de n indivíduos ou unidades experimentais. Este conjunto de $n \times p$ medidas origina uma matriz de dados **X** ($n \times p$):

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Para a obtençãodos componentes principais de uma forma geral, seja um conjunto de p variáveis $X_1, X_2, \ldots, X_p$ com médias $\mu_1, \mu_2, \ldots, \mu_p$ e variância $\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots, \sigma_p^2$, respectivamente. Estas variáveis não são independentes e portanto, possuem covariância entre a i-ésima e k-ésima variável definida por σ_{ik}, para $i \neq k = 1,2, \ldots, p$. Então as *p*variáveis podem ser expressas na forma vetorial por: $\mathbf{X} = [X_1, X_2, \ldots, X_p]'$, com vetor de médias $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_p]'$ e matriz de covariância**Σ**.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1p}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \cdots & \sigma_{nn}^2 \end{bmatrix}$$

Encontram-se os pares de autovalores e autovetores $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_P, e_P)$, em que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_P$, associados a **Σ** e então o i-ésimo componente principal é definido por (JOHNSON; WICHERN, 1998; HONGYU, 2015):

$$\boldsymbol{Z}_i = \boldsymbol{e}_i'\boldsymbol{X}$$
$$= e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_P$$

A variável $\boldsymbol{Z}_i$, é uma variável latente, ou seja, não é mensurada a partir do experimento ou levantamento amostral (JOHNSON; WICHERN, 1998). O objetivo é determiná-la a partir das p variáveis contidas no vetor **X**. A idéia é projetar os pontos coordenados originais em um plano maximizando a distância entre eles, o que equivale a maximizar a variabilidade da variável latente $\boldsymbol{Z}_i$. A variância de $\boldsymbol{Z}_i$ é dada por

$$Var(\boldsymbol{Z}_i) = Var(\boldsymbol{e}_i'\boldsymbol{X})$$
$$= \boldsymbol{e}_i' Var(\boldsymbol{X}) \boldsymbol{e}_i$$
$$= \boldsymbol{e}_i' \boldsymbol{\Sigma} \boldsymbol{e}_i$$

em que i = 1, . . . , p.

Utilizando a decomposição espectral da matriz $\boldsymbol{\Sigma}$, dada por$\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$, emque$\boldsymbol{P}$é a matriz composta pelos autovetores de $\boldsymbol{\Sigma}$em suas colunas e$\boldsymbol{\Lambda}$, a matriz diagonal de autovalores de$\boldsymbol{\Sigma}$, então, tem-seque

$$tr(\boldsymbol{\Sigma}) = tr(\boldsymbol{P\Lambda P'}) = tr(\boldsymbol{\Lambda P'P}) = tr(\boldsymbol{\Lambda I}) = tr(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} \lambda_i$$

e

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix}$$

E$tr(\boldsymbol{\Sigma})$ é dada pela soma dos elementos da diagonal:

$$tr(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i$$

Portanto, a variabilidade total contida nas variáveis originais é igual a variabilidade total contida nos componentes principais (JOHNSON; WICHERN, 1998). A contribuição de cada componente principal ($\boldsymbol{Z}_i$) é expressa em porcentagem, e a explicação individual de cada componente pode ser calculada, por exemplo, para k-ésimo componente principal a proporção da explicação é dada por:

$$C_k = \frac{Var(\boldsymbol{Z}_i)}{\sum_{i=1}^{p} Var(\boldsymbol{Z}_i)}.100 = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}.100 = \frac{\lambda_i}{tr(\boldsymbol{S})}.100$$

Pela proporção de explicação da variância total, que o modelo de *k* componentes principais é responsável, podemos determinar o número de componentes que deve-se reter. Em muitos casos, adota-se modelos que expliquem pelos menos 80% da variação total (JOHNSON; WICHERN, 1998).

Em geral escolhe-se o componente principal de maior importância (o primeiro componente principal) como sendo aquele de maior variância ($\max_i \lambda_i$), que explique o máximo de variabilidade dos dados, o segundo componente de maior importância, o que apresenta a segunda maior variância e assim sucessivamente, até o componente principal de menor importância (MANLY, 1986). Por outro lado, os últimos componentes principais serão responsáveis por direções que não estão associadas a muita variabilidade. Em outras palavras, esses últimos componentes principais identificarão relações lineares entre as variáveis originais próximo de constante (JOHNSON; WICHERN, 1998; ANDERSON, 2003; FERREIRA, 2011).

Outro método muito utilizado pelos pesquisadores é o critério de Kaiser (KAISER, 1958) para selecionar os componentes principais (CPs) que explicam a maior parte da variação dos dados. Este critério obtém CPs com valores próprios maiores do que a unidade ($\lambda_i > 1$), isto é, os principais componentes que explicam a maior parte da variação no conjunto de dados (SAVEGNAGO et al., 20011). O critério adotado por Jolliffe (1972, 1973) para descartar o número de CPs deve ser igual ao número de componentes cuja variância é inferior a 0,7 ($\lambda_i < 0{,}7$).

RESULTADOS E DISCUSSÕES

Com base nos resultados obtidos pela técnica dos componentes principais, os respectivos autovalores e porcentagens da variância explicada por cada um estão apresentados na Tabela 1. Os dois primeiros PCs foram responsáveis por **68.13**% da variação total, sobre a taxa de criminalidade de algumas cidades dos Estados Unidos, em que o PC1 foi responsável por **49,37**% e o segundo, PC2, por **18,76**% das variações dos dados.

Paiva et al., (2010), em estudo com 11 características de produção com aves de corte, verificaram que apenas três componentes principais eram suficientes para explicar 77% da variância total das características. Enquanto que Meira et al., (2013) em 13 características morfofuncionais de cavalos da raça Mangalarga Machador obtiveram 6 componentes principais com autovalores inferiores a 0,7 os quais explicaram 78,57%

da variação total das informações (FRAGA et al., 2015).

Para a determinação do número de componentes principais, verificou-se que como os dois primeiros CPs gerados a partir desta análise que tem autovalores > 1 ($\lambda_i > 1$) (Kaiser, 1958; FRAGA, et al., 2015) e foi responsável por 68,13% da variância total no conjunto de dados, os dois CPs foram retidos, com o auxílio do *screeplot* (Figura 1) e estão apresentados na Tabela 2. Portanto, dois primeiros componentes principais resumem efetivamente a variância amostral total e podem ser utilizados para o estudo do conjunto de dados.

Tabela 1:Componentes principais (CPs), autovalores (λ_i) e porcentagem da variância explicada e proporção acumulada (%) pelos componentes.

Componente Principal	Autovalores	Proporção	Proporção Acumulada (%)
PC1	3,45	49,37	49,37
PC2	1,31	18,76	68,13
PC3	0,97	13,89	82,03
PC4	0,58	8,36	90,40
CP5	0,39	5,64	96,04
CP6	0,17	2,44	98,49
CP7	0,10	1,50	100,00

Figura 1.O *screeplot* dos autovalores dos componentes principais.

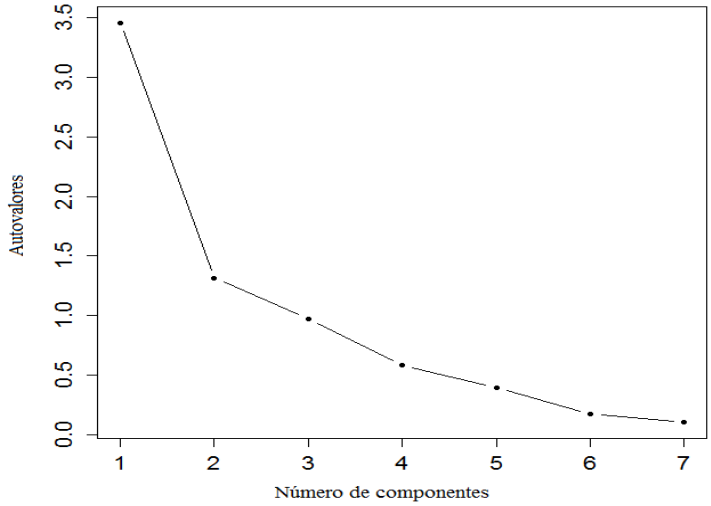


Tabela 2: Coeficientes de ponderação das características e seus coeficientes de correlações com os dois primeiros componentes principais.

Variável	Coeficiente de ponderação		Correlação	
	CP1	CP2	CP1	CP2
X_1	0,27	0,61	0,51	0,70
X_2	0,43	0,09	0,81	0,10
X_3	0,38	0,13	0,71	0,15
X_4	0,46	0,28	0,85	0,33
X_5	0,38	-0,39	0,72	-0,45
X_6	0,35	-0,59	0,65	-0,67
X_7	0,31	-0,10	0,59	-0,12

Como intuito de se entender a importância de cada variável na construção dos dois componentes foi calculado a correlação entre as variáveis originais e os componentes principais, apresentados também na Tabela 2.

Na Tabela 2, são apresentados, ainda, as correlações com os dois primeiros componentes principais e seus coeficientes de ponderação de cada característica.

Com a seleção de dois componentes principais, a redução da dimensão de 7 variáveis originais para 2 componentes principais é bastante razoável. Portanto decidiu-se utilizar unicamente os dois primeiros componentes principais para a composição das equações 1 e 2.

$$CP1 = 0.27X_1 + 0.43X_2 + 0.38X_3 + 0.46X_4 + 0.39\ X_5 + 0.35\ X_6 + 0.31X_7 \quad (1)$$

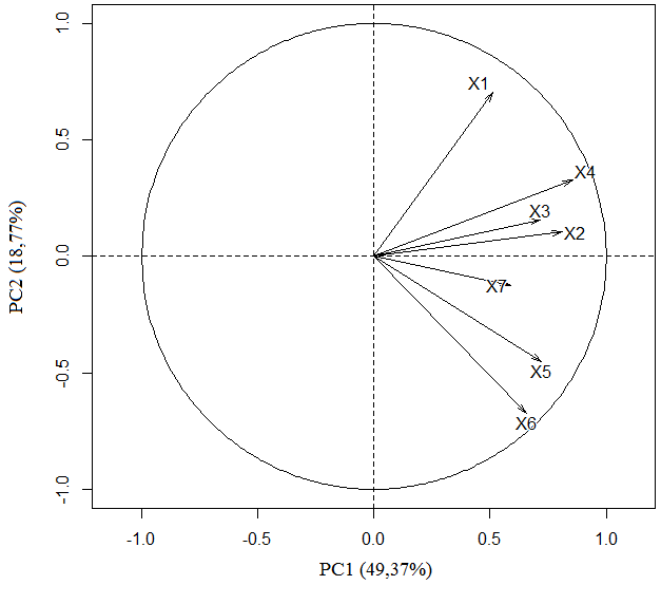
$$CP2 = 0.61X_1 + 0.09X_2 + 0.14X_3 + 0.29X_4 - 0.39X_5 - 0.59\ X_6 - 0.11X_7 \quad (2)$$

De acordo com a equação (1) e a Tabela 2, no primeiro componente principal destacaram-se as variáveis X_2 (Estupro) e X_4 (Assalto) e neste caso pode-se chamá-lo de componente de crimes de estupro e assalto. E de acordo com a equação (2) e Tabela 2, no segundo componente principal ficou evidente o contraste entre X_1 (assassinato) e X_6 (Pequenos furtos), podendo ser chamado de componente contraste de crimes de assassinato com (e) pequenos furtos.

As variáveis X_2 e X_4 apresentaram contribuições similares para o CP1, isto foi verificado pelas variáveis que têm vetor de maior comprimento e que foram mais próximas ao eixo CP1, mostrado na Figura 2. Existem correlações altas entre as variáveis X_2, X_3 e X_4, pois formaram ângulos agudos entres as variáveis, também as variáveis X_5 e X_6. Não existe correlação entre as variáveis X_1 e X_6, pois forma um ângulo próximo de 90 graus, como mostrado na Figura 2.

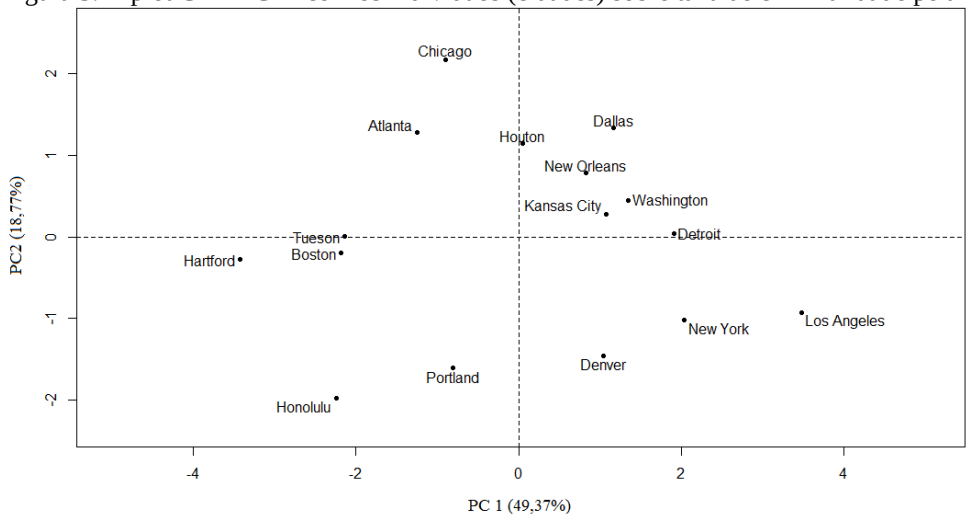
ACP foi usada para reduzir as dimensões das variáveis originais sem perda de informação. Por definição, a correlação entre os principais componentes é zero, isto é, a variação explicada em CP1 é independente da variação explicada em CP2 e assim por diante. Isto implica que para qualquer componente principal não vai causar uma resposta correlacionada em termos de outros componentes principais, isto é, eles são ortogonais (SAVEGNAGO et. al., 2011; FRAGA et al., 2015).

Figura 2. Biplot CP1 × CP2 sobre as variáveis (taxa de criminalidade) em algumas cidades dos Estados Unidos pela ACP



E&S - Engineering and Science ISSN: 2358-5390 DOI: 10.18607/ES20165053 Volume 1, Edição 5

Figura 3. Biplot CP1 × CP2 com os indivíduos (cidades) sobre taxa de criminalidade pela ACP.



Como os dois primeiros componentes principais explicaram 68,13% da variação total dos dados (Figura 3), de acordo com Rencher (2002), pelo menos 70% da variância total devem ser explicadas pelos primeiros e o segundo componentes principais.

Analisando a Figura 3, que é o BiplotCP1xCP2 com as cidades sobre taxas de criminalidade e as equações (1) e (2), pode-se concluir que, de acordo com os dados de criminalidade das cidades dos Estados Unidos e com a ACP, *New York*, *Los Angeles*, *Detroit* e *Washington* possuem maiores ocorrências de taxa de criminalidade do país e principalmente sobre maiores números de crimes de estupro e assalto pela CP1. *Hartford*, *Honoluiu*, *Boston* e *Tueson* foram às cidades que apresentaram menores números de ocorrências de taxa de criminalidade, principalmente sobre estupro e assalto.

E pela CP2, conclui-se que cidades como *Chicago*, *Atlanta*, *Dallas* e *Houton* são as que apresentaram maiores números de ocorrências de taxa criminalidade sobre assassinato e menores números de ocorrências sobre pequenos furtos. As cidades como *Honolulu*, *Portland* e *Denver* apresentaram maiores números de ocorrências sobre pequenos furtos e menores números sobre assassinato.

CONCLUSÃO

Tendo em vista os resultados obtidos, a análise de componentes principais se mostrou efetiva e permitiu a retirada ou descarte de cinco variáveis que apresentaram baixa variabilidade ou foram redundantes por estarem correlacionadas com as de maior importância para dois componentes principais. Assim, um menor número de variáveis foram necessárias para explicar a variação total resultando em economia de tempo e de recursos em futuros trabalhos que utilizarão essa mesma base de dados, sem perda significativa de informação.

Um dos objetivos da ACP, neste caso, foi atingido, pois um número relativamente pequeno de componentes foi extraído (CP1 e CP2) com a capacidade de explicar a maior variabilidade nos dados originais (68,13%).

REFERÊNCIAS

ANDERSON, T.W. **An introduction to multivariate statistical analysis**.New York: Wiley, 6 ed. 2003. 374p.

HONGYU, K. **Distribuição empírica dos autovalores associados à matriz de interação dos modelos AMMI pelo método bootstrapnão-paramétrico**. 2012. 104p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola

Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2012.

HONGYU, K.**Comparação do GGE-biplot ponderado e AMMI-ponderado com outros modelos de interação genótipo × ambiente**. 2015. 155p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2015.

HOTELLING, H. Analysis of a complex of statistical variables into principal components.**The Journal Educational Psychology**, Cambridge, v.24, p.498-520, 1933.

HOTELLING, H. Simplifield calculation of principal components.**Psychometrika**, Williamsburg, v.1, p.27-35, 1936.

FERREIRA, D.F. **Estatística Multivariada**.Lavras: UFLA, 2011. 675p.

FRAGA, A.B.; SILVA, F.L.; HONGYU, K.; SANTOS, D.D.S.; MURPHY, T.W.; LOPES, F.B. Multivariate analysis to evaluate genetic groups and productiontraits of crossbred Holstein × Zebu cows. **Trop Anim Health Prod.**p. 1-6. 2015

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**.Madison: Prentice Hall International, 1998. 816p.

JOLLIFFE, I.T. Discarding Variables in a Principal Component Analysis. I: Artificial **Journal of the Royal Statistical Society**. Series C (Applied Statistics), Vol. 21, No. 2 p. 160-173, 1972.

JOLLIFFE, I.T. Discarding Variables in a Principal Component Analysis. II: Real Data. **Journal of the Royal Statistical Society**, v. 22, n. 1, p. 21–31, 1973.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis.**Psychometrika**, v. 23, n. 3.p. 187-200, 1958.

MANLY, B. F. J.**Multivariate statistical methods**.New York, Chapman and Hall, 1986.159 p.

MEIRA, C. T.; PEREIRA, I. G.; FARAH, M. M.;PIRES, A. V.; GARCIA, D. A.; CRUZ, V. A. R. Seleção de características morfofuncionais de cavalos da raça Manga larga Marchador por meio da análise decomponentes principais, **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, 65, 1843–1848. 2013.

PAIVA, A.L.C.; TEIXEIRA, R.B.; YAMAKI, M.; MENEZES, G.R.O.; Leite, C. D.S.; TORRES, R.A. Análise de componentes principais emcaracterísticas de produção de aves de postura, **Revista Brasileira de Zootecnia**, 39, 285–288.2010.

R DEVELOPMENT CORE TEAM. R: a language and environment for statistical computing. Vienna: R Foundation for StatisticalComputing, Vienna, 2014.
REGAZZI, A.J. Análise multivariada, notas de aula INF 766, Departamento de Informática da Universidade Federal de Viçosa, v.2, 2000.

RENCHER, A.C. **Methods of Multivariate Analysis**.A JOHN WILEY & SONS, INC. PUBLICATION.p.727. 2ed. 2002.

SAVEGNAGO, R.P., CAETANO, S.L., RAMOS, S.B., NASCIMENTO, G.B., SCHMIDT, G.S., LEDUR, M.C. MUNARI, D.P. Estimates ofgenetic parameters, and cluster and principal components analysesof breeding values related to egg production traits in a White Leghorn population, **Poultry Science**, 90, p.2174-2188. 2011.

SAS Institute Inc. 2008.SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc. 1st electronic book