Revista de Estudos e Pesquisas sobre Ensino Tecnológico

ISSN: 2446-774X

Ensino de técnica de estatística multivariada para alunos de cursos de Engenharia

Teaching multivariate statistical techniques to students of engineering courses

Fernanda Santana Carvalho https://orcid.org/0000-0001-6215-6930
Departamento de Engenharia de Transportes - Escola Politécnica, Universidade de São Paulo
Faculdade de Ciências e Tecnologia, Universidade Federal de Goiás
e-mail – carvalhofernanda@ufg.br

Cláudia A. Soares Machado https://orcid.org/0000-0001-6276-4645
Departamento de Engenharia de Transportes - Escola Politécnica, Universidade de São Paulo
e-mail - claudia.machado@usp.br

José Alberto Quintanilha https://orcid.org/0000-0003-3261-7825
Instituto de Energia e Ambiente, Universidade de São Paulo
e-mail - jaquinta@usp.br

Resumo

A Análise de Componentes Principais (ACP) é uma ferramenta padrão em análise multivariada aplicada para examinar dados multidimensionais e amplamente utilizada em diferentes áreas do conhecimento. Destaca-se devido ao aumento no tamanho das bases de dados e surgimento do *Big Data*. Seu ensino em cursos de engenharia se torna importante, principalmente quando aliado ao uso de *softwares* livres e gratuitos, como o "RStudio". Este artigo tem o objetivo de apresentar as experiências produzidas a partir de um estudo de caso realizado com estudantes do curso de engenharia, onde o exemplo utilizado para a aplicação da ACP são tabelas relativas às mortes decorrentes de acidentes de trânsito, em diferentes países do mundo. O estudo evidenciou a empregabilidade da ferramenta ACP em aulas ministradas na graduação em engenharia com o intuito de mostrar, de maneira simplificada, como iniciar uma análise com múltiplas variáveis utilizando o "RStudio", *software* de fácil acesso aos alunos, caracterizado como um ambiente computacional e linguagem de programação para análise estatística e visualização gráfica de dados. Demonstrou-se a utilidade da ACP como ferramenta introdutória para análises multivariadas no ensino em cursos de engenharia e, por meio de sua utilização os estudantes puderam observar padrões e correlações existentes no estudo de caso. Pode-se concluir que a metodologia adotada, que consistiu de aula expositiva seguida de prática guiada por vídeos tutoriais, atingiu seu objetivo, pois os alunos puderam replicar seu aprendizado de forma autônoma e independente, sem a necessidade de dispender recursos financeiros próprios ou da universidade.

Palavras-chave: Ensino Superior. Tecnologia e Didática. Estatística.

Abstract

PCA (principal components analysis) is a standard applied multivariate analysis tool to survey multidimensional data, and it is largely used in different fields of knowledge. It is gaining prominence with the increase in the size of databases and the emergence of Big Data. Teaching PCA in engineering courses is becoming important, especially when allied to free software, such as RStudio

(easily available to students and characterized by a computer environment and programming language for statistical analysis and graphic data visualization). This article presents a case study of engineering students who applied PCA to data on traffic accident deaths in different parts of the world. The study brings the usefulness of PCA in undergraduate engineering courses, showing in a simplified manner, how to start a multivariate analysis using RStudio. The usefulness of PCA as an introductory tool for teaching multivariate analysis in engineering courses was demonstrated, as students could observe patterns and correlations in the case study. The methodology – lectures followed by practical guided video tutorials – achieved its objective, since students could learn freely and independently, avoiding the use of the university financial resources or their own.

Introdução

A Análise de Componentes Principais (ACP) é uma ferramenta padrão em análise multivariada aplicada para examinar dados multidimensionais e, por meio dela, torna-se possível entender a estrutura desses dados e observar correlações, bem como agrupamentos das variáveis observadas. Aplicada em diferentes áreas do conhecimento, essa ferramenta ganhou ainda mais visibilidade na atualidade devido ao custo e à facilidade de aquisição de dados nunca terem sido tão baixos, fazendo-os serem coletados por mais pessoas, durante períodos maiores e provenientes de inúmeras fontes, além de ter dado início às bases de dados popularmente conhecidas como *Big Data* (KLAŠNJA-MILIĆEVIĆ; IVANOVIĆ; BUDIMAC, 2017).

No campo da engenharia, uma vez que os profissionais têm a necessidade de lidar cada vez mais com esse grande volume de informações, torna-se fundamental o ensino de ferramentas de manipulação, tratamento e análise de dados já durante sua graduação. Várias são as fontes de dados com possível aplicação da ACP no ensino de engenharia. Dentre elas, encontra-se o estudo da segurança viária, área da Engenharia de Transportes que usualmente trabalha com informações obtidas nos acidentes de trânsito: data, horário, local, número de pessoas envolvidas, número de veículos envolvidos, dentre outras informações.

A segurança viária é um assunto que tem ganhado protagonismo mundial por ter se tornando um problema de saúde pública. Isso acontece pelo fato de que, segundo a Organização Mundial da Saúde (OMS, 2018), quase 1,35 milhões de pessoas morrem por ano devido a acidentes de trânsito. A fim de um efetivo enfrentamento do problema, a adoção de medidas mitigadoras tem que partir do estudo das causas do problema. Vários autores, dentre eles Gold (1999), caracterizam o acidente de tráfego como resultado da interação de um conjunto de fatores, dentre eles: o fator humano, o veículo, a via e as condições do ambiente no momento do acidente, além de aspectos institucionais e sociais.

Nesse sentido, Wegman (2016) afirma que as medidas para a redução de acidentes envolvem também esforços sociais como a implementação de inspeções veiculares e o melhoramento do comportamento humano pela legislação e através de campanhas educativas. Essas campanhas envolvem o ensino da segurança viária, que deve abranger de forma precisa os jovens, uma vez que, segundo a Organização Mundial da Saúde, são eles as maiores vítimas dos acidentes no trânsito, sendo essa a principal causa de morte de jovens entre 5 e 29 anos (OMS, 2018). Dessa forma, o ensino da segurança viária, presente em alguns cursos de Engenharia Civil e de Transportes no Brasil, apresenta, também, uma função social.

O presente trabalho visa mostrar a empregabilidade da ferramenta ACP em aulas ministradas na graduação em engenharia com o intuito de mostrar, de maneira simplificada, como iniciar uma análise com múltiplas variáveis utilizando um

software livre. O exemplo utilizado são duas tabelas que apresentam a quantidade de mortes por acidentes trânsito em diferentes países do mundo, a fim de evidenciar padrões e correlações no *ranking* de mortes no trânsito, debatendo também a segurança viária em nível mundial e apresentando aos alunos uma ferramenta que viabiliza discussões sobre a urgência de medidas mitigadoras efetivas para que sejam evitadas novas mortes. O *software* livre utilizado foi o "RStudio", um *software* gratuito e de fácil acesso aos alunos, caracterizado como um ambiente computacional e linguagem de programação para análise estatística e visualização gráfica de dados.

Processos de Ensino e Aprendizagem em Engenharia

Segundo Aranha (1996), a interdisciplinaridade no conteúdo trabalhado por professores é permitida a partir da busca de novas metodologias pedagógicas. Para a autora, essa busca deve prevalecer tanto na pesquisa quanto na Educação.

Siqueira *et al.* (2012), citando Delors (1998), apontam que a prática pedagógica deve se preocupar em desenvolver quatro aprendizagens fundamentais: aprender a saber, aprender a fazer, aprender a viver e aprender a ser. No caso apresentado neste artigo, buscaram-se duas dessas aprendizagens. A primeira, "aprender a saber", segundo Delors (1998), no sentido de reforço de conteúdos de disciplinas anteriores como Álgebra Linear, Cálculo Numérico e Estatística, e a segunda, "aprender a fazer" no sentido de Delors (1998), pela execução de processamento computacional orientado.

Segundo Bartholomew (2010), o principal objetivo da ACP é tornar a estrutura de dados multivariados mais clara pela redução de sua dimensionalidade. Segundo ele, o processo se dá pela procura da combinação linear das variáveis responsável pela máxima variação total possível dos dados. Para a segunda combinação, tal processo se repete com o restante dos dados e assim por diante. Ao final, tendo sido contabilizadas um número pequeno de componentes, substitui-se o conjunto original de variáveis pelas mesmas.

Essa descrição resume a abordagem adotada neste artigo e refere-se a Estatística (dados multivariados) e Álgebra Linear (combinação linear). No caso da Estatística, Ara (2006) destaca a dificuldade de ensinar e aprender seus conceitos para cursos de engenharia, associando esse fato à falta de reconhecimento prévio e adequado de fenômenos aleatórios.

Westfall, Arias e Fulton (2017) afirmam que introduzir componentes principais para os alunos é difícil, especialmente para os alunos das ciências sociais e comportamentais. Gajewski *et al.* (2014) discutem as dificuldades de ensino de análise multivariada para não estatísticos e usam o ponto de vista do estudo de caso, argumentando que essa abordagem tem uma vantagem pedagógica por revisar materiais de um curso anterior, assim como o material desenvolvido e enfatizado neste texto.

O manuscrito de Souza e Poppi (2011) tem como objetivo apresentar os conceitos básicos e a aplicação prática da ACP como tutorial, para alunos iniciantes, graduação e pós-graduação. Como exemplo prático é mostrada a análise exploratória de óleos vegetais comestíveis por espectroscopia de infravermelho médio. De forma semelhante, Valderrama *et al.* (2016) propuseram um experimento

didático envolvendo ACP, considerado um método quimiométrico que permite a extração de informações químicas, que de outra forma seriam impossíveis de determinar.

Análise de Componentes Principais

Para Jolliffe (2002) a Análise de Componentes Principais (ACP) é possivelmente a mais antiga e mais conhecida técnica de análise multivariada, sendo introduzida pela primeira vez por Pearson (1901) e desenvolvida posteriormente e independente por Hotelling (1933), e conhecida também na área de processamento de sinais como Transformada de Karhunen-Loève (DONY, 2001).

Uma ACP pode ser definida como uma ferramenta padrão em análise multivariada para examinar dados multidimensionais, ou seja, com muitas variáveis. Além disso, tem a capacidade de revelar padrões entre objetos ou variáveis, que não seriam aparentes em uma análise univariada.

Dessa forma, a ACP tem dois objetivos principais: o primeiro é entender a estrutura de um grande conjunto de variáveis e o segundo é reduzir um conjunto multidimensional de dados, com seus padrões e relações, para um tamanho gerenciável, retendo o máximo possível de informações originais e com perda de informação conhecida (PAPADIMITRIOU, YANNIS, 2013).

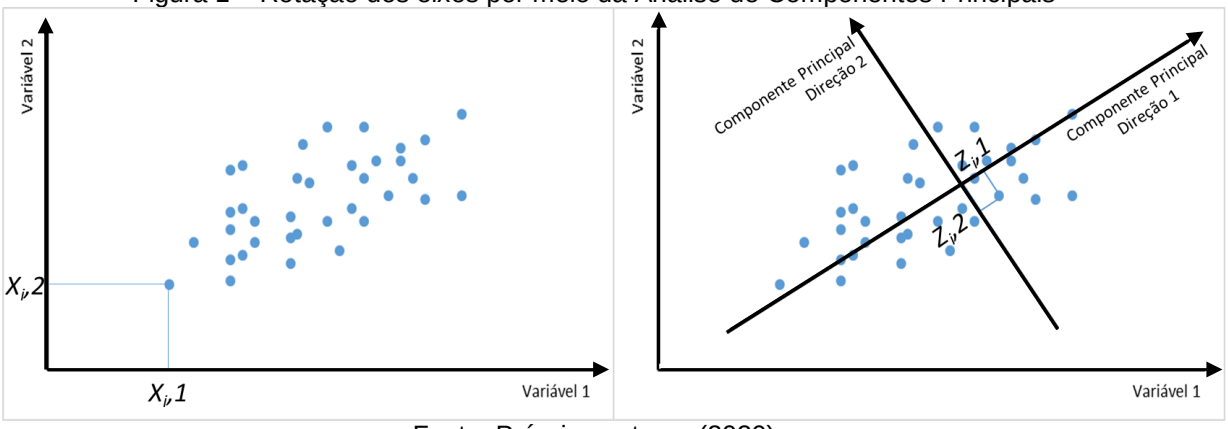
Pitombo e Gomes (2014) incluem que a partir da estrutura de dependência entre as variáveis do conjunto de dados inicial, a ACP possibilita a criação de um conjunto menor de variáveis, as denominadas componentes. As autoras explicam ainda que é possível saber até que ponto cada uma das componentes geradas está associada a cada variável do conjunto de dados e quanto esses componentes explicam a variabilidade do conjunto de dados originais.

Devido a essas características, a ACP tem usos diversos e se aplica, principalmente: em análises exploratórias de dados, na detecção de *outliers* (valor atípico que apresenta um grande afastamento dos demais da série), na identificação de *clusters* (agrupamentos, regionalização), na redução de variáveis (pré-processamento de dados, modelagem multivariada), na compressão de dados com perdas conhecidas, na análise da variabilidade no espaço e no tempo de um conjunto de variáveis, na nova interpretação dos dados em termos dos principais componentes da variabilidade, e na previsão de variáveis possibilitada pela busca de relações entre elas.

A ACP se vale da ideia de que se existe uma correlação diferente de zero entre as variáveis do conjunto de dados é possível determinar uma descrição mais compacta dos dados. A correlação entre variáveis significa redundância nos dados, ou seja, se duas variáveis são perfeitamente correlacionadas, então, uma delas é redundante porque se sabemos X, deduzimos Y.

Dessa forma, a ACP explora essa redundância e reduz o conjunto de dados correlacionados (valores das n variáveis $\{X_1, X_2, ..., X_n\}$) para um conjunto de dados contendo menos e novas variáveis por rotação dos eixos, conforme Figura 1, de forma que as novas variáveis sejam combinações lineares não correlacionadas das originais (linearmente independentes), resumindo grande parte das variáveis originais com perda conhecida de informações.

Figura 1 – Rotação dos eixos por meio da Análise de Componentes Principais



Fonte: Próprios autores (2020)

Na Figura 1, vê-se que no gráfico da esquerda estão lançados os valores observados de duas variáveis (variável 1 no eixo das abscissas e variável 2 no eixo das ordenadas) e percebe-se que há um alinhamento dos pontos numa mesma direção, direção esta que indica a existência de maior variabilidade (varia numa amplitude maior). Essa será a direção do novo eixo mostrado no gráfico da direita com o nome de Componente Principal Direção 1. Ortogonalmente a esse eixo e na intersecção com a sua origem, está a Componente Principal Direção 2. Isso, geometricamente, corresponde a uma translação (da origem inicial para a nova origem dos componentes) e uma rotação (na direção de maior variabilidade) dos eixos iniciais variável 1 e variável 2.

Sendo assim, na ACP, a partir de *p* variáveis num espaço *p* dimensional (chamado de espaço de atributos ou de características), faz-se uma rotação dos eixos de modo a se obter novos *p* eixos que possuem as seguintes propriedades: são ortogonais entre si, contém todo o montante de informação das *p* variáveis originais, são ordenados de tal forma que o eixo da Componente Principal na Direção 1 tenha a maior proporção da variância (variabilidade) total das variáveis originais, o eixo da Componente Principal na Direção 2 tenha a segunda maior proporção da variância e o eixo da Componente Principal na Direção *p* tenha a menor proporção de variância do montante inicial. Além disso, como dito anteriormente, a covariância e a correlação entre cada par dos novos eixos principais é zero, ou seja, os eixos principais não são correlacionados.

De acordo com Lindner *et al.* (2016), a formulação matemática da ACP é baseada em uma matriz de variância-covariância (matriz *S*) ou matriz de correlação linear (matriz *R*) conforme descrito a seguir. A matriz de variância-covariância, denota a dispersão dos dados e é usada para dados numa mesma escala de medida, enquanto a matriz de correlação considera dados medidos em diferentes escalas.

Dada uma matriz *S* (*n* × *n*), onde *n* é o número de variáveis, as componentes principais são dadas pelo cálculo autovetores (v) e autovalores λ da matriz S.

Os autovetores (v) são calculados em função dos autovalores (λ) da matriz *S*. *I* é a matriz de identidade, e os autovalores (λ) da matriz S são escalares que satisfazem a equação característica Eq. (1):

$$\left|S - \lambda I\right| = 0 \qquad (1)$$

Cada valor próprio é associado a um vetor próprio, que pode ser obtido na Eq. (2):

$$\left(S-\lambda I\right)v=0 \quad (2)$$

No caso geral, a matriz de autovalores é diagonal, onde o número de autovalores é equivalente a uma matriz quadrada ($n \times n$). Um novo conjunto de variáveis pode ser derivado multiplicando os autovetores e os vetores dos valores originais. Portanto, uma matriz quadrada A é composta usando autovetores como colunas da matriz. O novo conjunto de variáveis (W) é uma combinação linear das variáveis originais, derivada da Eq. (3):

$$W=XA \quad (3)$$

Onde a matriz *A* compreende os autovetores e *X* é o vetor de dados original. As componentes principais são selecionadas verificando a fração da variação que é explicada por um componente específico. Quanto maior a sua proporção, mais relevante é a componente para a análise (LINDNER *et al*., 2016).

Metodologia

A aula foi conduzida por um professor, com auxílio de duas monitoras, uma da pós-graduação e outra da graduação que já havia cursado a disciplina. A fim de facilitar o entendimento do conteúdo, a aula foi dividida em duas partes, sendo a primeira teórica e a segunda prática.

Na parte teórica a Análise de Componentes Principais foi definida matematicamente, pelo uso de *slides* expositivos, e demonstrados seus possíveis empregos, bem como seus benefícios pela apresentação de artigos científicos.

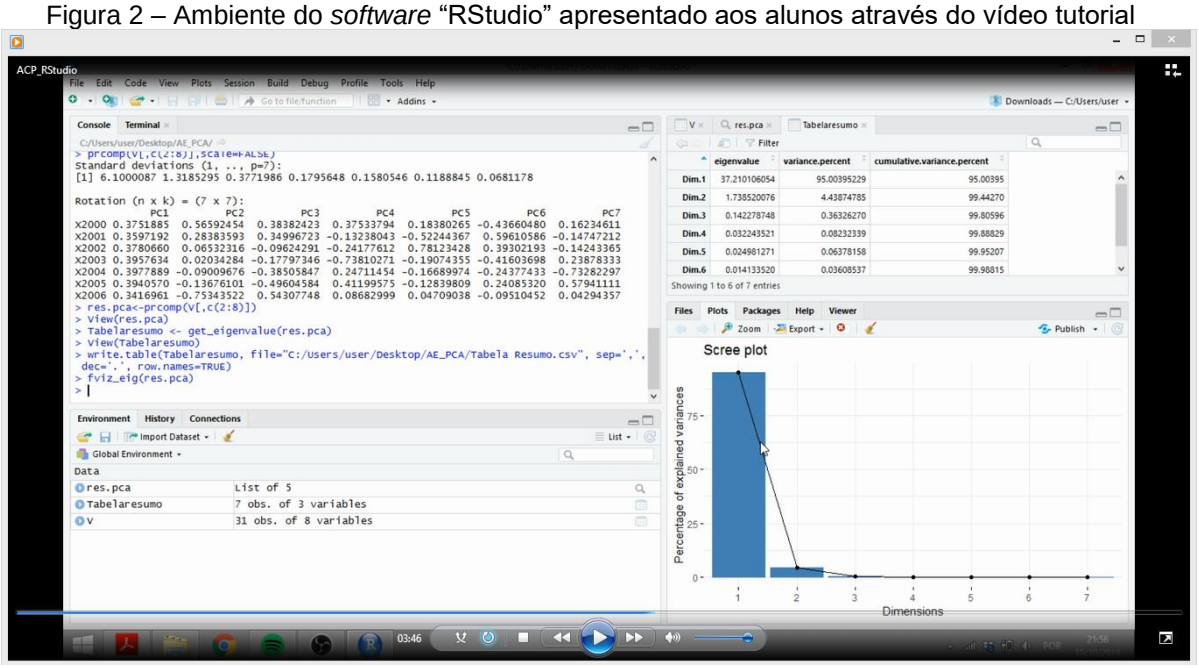
Por sua vez, a parte prática iniciou-se com a apresentação do estudo de caso por meio de *slides.* Na sequência, foi apresentado um tutorial em forma de vídeo, a fim de facilitar o entendimento do emprego do *software* “RStudio”, conforme Figura 2, e no formato de *script*. *Script* é o nome dado ao conjunto de etapas a serem seguidas, bem como, ao conjunto de comandos a serem inseridos no *software*. Dessa forma, os estudantes acompanhavam e realizavam as etapas/operações em tempo real ao seguir os vídeos e o material escrito com o auxílio das monitoras.

O conjunto de dados utilizado no estudo de caso da parte prática foi o *ranking* de acidentes por país obtido por Lopes (2012) e apresentado em mortes por 10 mil veículos, conforme a Tabela 1. Os dados observados durante os anos de 2000 a 2006 abrangem 31 países, incluindo o Brasil, que se encontrava na 27ª colocação, sendo a primeira colocação aquela com menor frequência de acidentes.

A primeira atividade da parte prática consistiu na adequação da planilha de dados e na substituição de valores não observados, apresentados como “-” na Tabela 1, sendo explicado aos estudantes o efeito dessa prática na obtenção dos resultados. Em seguida, ACP foi realizada de duas formas diferentes: (1) segundo o posicionamento dos países e (2) segundo o ano observado. Cada gráfico produzido era salvo conforme determinado no *script*.

A ideia foi a proposição de uma atividade interativa com os alunos, por meio de uma aplicação prática dos conceitos abstratos apresentados aos estudantes nas disciplinas teóricas de Álgebra Linear e Estatística para engenharia.

Figura 2 – Ambiente do *software* “RStudio” apresentado aos alunos através do vídeo tutorial



Fonte: Próprios autores (2020)

Buscando-se induzir a reflexão dos estudantes e firmar os conhecimentos passados em aula, foi apresentada a planilha de *ranking* de acidentes, desta vez em acidentes por 100 mil habitantes (LOPES, 2012 - não apresentada no texto), a fim de que as análises fossem refeitas e os resultados discutidos individualmente por cada um, tanto geograficamente e economicamente, em relação aos países, quanto numericamente em relação aos dados.

Tabela 1 – *Ranking* de acidentes por 10 mil veículos

País	2000	2001	2002	2003	2004	2005	2006
Suíça	1,5	1,4	1,3	1,3	1,2	1,0	0,9
Noruega	1,5	1,2	1,3	1,2	1,0	0,9	0,9
Suécia	1,3	1,3	1,3	1,2	1,1	1,0	1,0
Holanda	1,5	1,4	1,3	1,3	1,0	0,9	1,0
Reino Unido	1,3	1,2	1,2	1,2	1,0	1,0	1,0
Alemanha	1,6	1,5	1,4	1,4	1,2	1,1	1,0
Japão	1,6	1,5	1,4	1,3	1,2	1,2	1,1
Luxemburgo	2,5	2,2	1,9	1,6	1,5	1,4	1,1
Austrália	1,5	1,4	1,4	1,3	1,2	1,2	1,2
Finlândia	1,6	1,7	1,6	1,4	1,4	1,3	1,2
Dinamarca	2,2	1,9	2,0	1,9	1,6	1,4	1,2
Nova Zelândia	1,8	1,7	1,5	1,6	1,5	1,3	1,3
França	2,3	2,2	2,1	1,6	1,4	1,5	1,3
Itália	1,8	1,8	1,8	1,6	1,4	1,4	1,3
Portugal	2,7	2,3	2,2	2	1,6	1,5	1,3
Islândia	1,8	1,2	1,5	1,1	1,1	0,9	1,4
Espanha	2,6	2,4	2,3	2,3	1,9	1,6	1,6
Canadá	1,7	1,6	1,6	1,5	1,5	1,6	-
Áustria	2,2	2,1	2,2	2,1	2	1,7	1,6
Bélgica	2,8	2,8	2,5	2,2	2,1	2,0	1,9
Estados Unidos	2,0	2,0	1,9	1,9	1,9	1,9	1,9
Irlanda	2,6	2,4	2,2	1,8	2,1	1,9	2,0
República Tcheca	3,9	3,4	3,6	3,5	3,3	2,9	2,4
Grécia	4,8	4,1	3,5	3,2	3,2	3,0	2,9
Polônia	5,3	4,4	4,4	4,1	4,0	3,8	3,3
Hungria	4,4	4,3	4,7	4,1	3,8	3,7	3,8
BRASIL	6,8	6,3	6,2	6,2	6,5	6,2	4,3
Eslováquia	4,6	4,4	4,3	4,3	4,5	4,5	4,5
Coréia	9,2	7,0	6,2	5,9	5,2	4,9	-
Turquia	4,7	4,8	6,6	6,4	7,1	7,3	7,3
Rússia	11,7	11,7	12,0	12,8	12,4	12,2	11,8

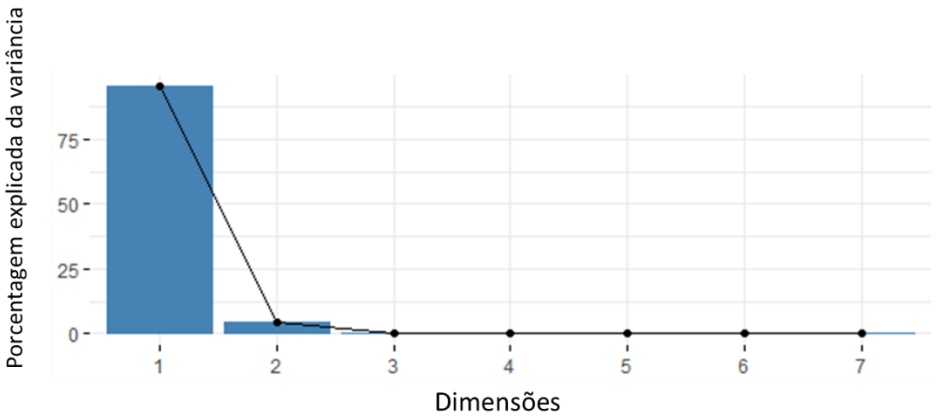
Fonte: Lopes (2012)

Resultados

Seguindo os tutoriais e o *script*, os alunos puderam produzir suas próprias análises, sendo a primeira delas referente ao número de acidentes por 10 mil veículos por ano por país. Essa análise resultou em 7 componentes principais devido ao fato de existirem 7 variáveis para cada país, sendo elas o número de acidentes por cada um dos 7 anos da base de dados (2000 a 2006).

No primeiro gráfico presente na Figura 3, observa-se que a primeira componente principal pode explicar quase 100% da variância dos dados. Na Tabela 2 esse fato se confirma com dados mais precisos para todas as 7 dimensões de componentes principais.

Figura 3 – Gráfico da porcentagem explicada da variância dos dados em cada dimensão das Componentes Principais



Fonte: Próprios autores (2020)

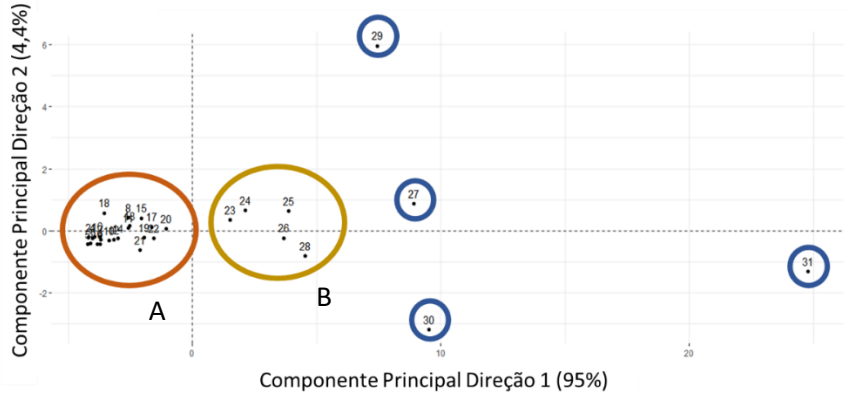
Tabela 2 – Autovalores de cada componente principal e a porcentagem explicada da variância em cada uma

	Autovalor	Variância Percentual	Variância Percentual Acumulada
Componente Principal 1	37,21	95,00	95,00
Componente Principal 2	1,74	4,44	99,44
Componente Principal 3	0,14	0,36	99,80
Componente Principal 4	0,03	0,08	99,88
Componente Principal 5	0,02	0,07	99,95
Componente Principal 6	0,01	0,04	99,99
Componente Principal 7	0,00	0,01	100,00

Fonte: Próprios autores (2020)

No próximo passo do *script*, os alunos deveriam produzir o gráfico presente na Figura 4, que apresenta as duas primeiras dimensões das componentes principais, de forma a visualizarem e discutirem os agrupamentos de países presentes no mesmo.

Figura 4 – Apresentação gráfica das duas primeiras dimensões da ACP em relação ao número de acidentes por 10 mil veículos por país



Fonte: Próprios autores (2020)

Como resultado, os alunos puderam concluir que houve a formação de três aglomerações segundo o número de acidentes: o grupo A, em laranja, continha países da Europa, América do Norte e Oceania; o grupo B em amarelo, abarcava apenas países europeus; já os últimos países em azul eram pontos não agrupados contendo os países da borda entre Europa e Ásia e o Brasil. A fim de compreender tais agrupamentos os alunos foram convidados a determinar para cada país a somatória do número de acidentes, a média e a variância dos mesmos mediante os anos, conforme a Tabela 3.

Tabela 3 – Tabela exploratória dos dados de acidentes por país

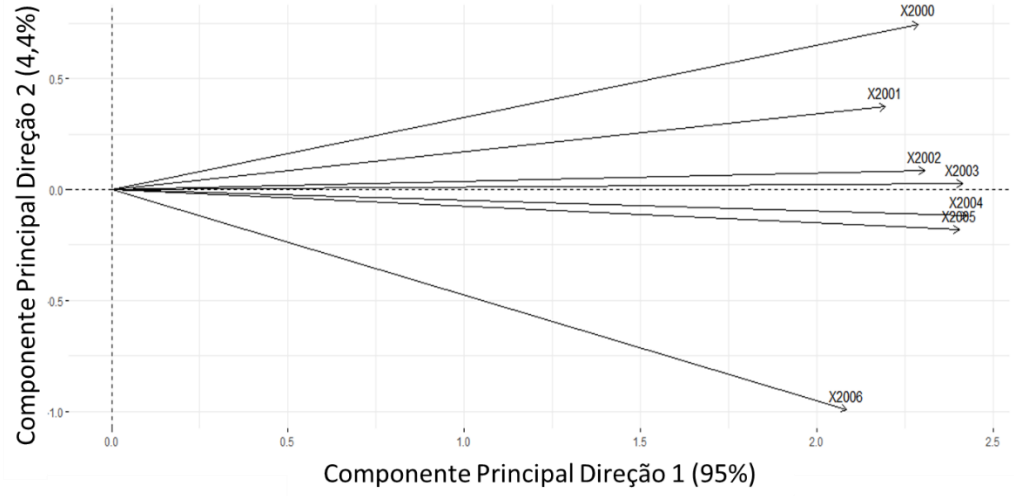
País	Total	Média	Desvio Padrão	Agrupamento
Suíça	8,60	1,23	0,21	A
Noruega	8,00	1,14	0,22	A
Suécia	8,20	1,17	0,14	A
Holanda	8,40	1,20	0,23	A
Reino Unido	7,90	1,13	0,13	A
Alemanha	9,20	1,31	0,22	A
Japão	9,30	1,33	0,18	A
Luxemburgo	12,20	1,74	0,49	A
Austrália	9,20	1,31	0,12	A
Finlândia	10,20	1,46	0,18	A
Dinamarca	12,20	1,74	0,36	A
Nova Zelândia	10,70	1,53	0,19	A
França	12,40	1,77	0,42	A
Itália	11,10	1,59	0,22	A
Portugal	13,60	1,94	0,50	A
Islândia	9,00	1,29	0,30	A
Espanha	14,70	2,10	0,40	A
Canadá	9,50	1,36	0,60	A
Áustria	13,90	1,99	0,24	A
Bélgica	16,30	2,33	0,37	A
Estados Unidos	13,50	1,93	0,05	A
Irlanda	15,00	2,14	0,28	A
República Tcheca	23,00	3,29	0,49	B
Grécia	24,70	3,53	0,69	B
Polônia	29,30	4,19	0,62	B
Hungria	28,80	4,11	0,37	B
Brasil	42,50	6,07	0,81	Sem agrupamento
Eslováquia	31,10	4,44	0,11	B
Coréia	38,40	5,49	2,81	Sem agrupamento
Turquia	44,20	6,31	1,12	Sem agrupamento
Rússia	84,60	12,09	0,41	Sem agrupamento

Fonte: Próprios autores (2020)

Para agrupamentos de países, por exemplo, observou-se que no grupo A encontravam-se aqueles considerados como “países desenvolvidos” e que apresentavam a média de acidentes por ano entre 1,14 e 2,14, considerada baixa em comparação aos demais grupos. Para o grupo B, observou-se com os alunos que, apesar de serem países europeus, os mesmos não eram considerados como “potências econômicas” e apresentavam uma média de acidentes por ano entre 3,29 e 4,44. Por fim, os países que não se agruparam apresentavam média superior a 5,49, sendo a do Brasil 6,07 e chegando a um máximo de 12,09 para a Rússia.

Na segunda fase da aula prática, a ACP deveria ser refeita para o número de acidentes por ano observado, não mais por país. O resultado produzido está apresentado na Figura 5. Nessa aplicação, ficou evidente para os alunos o agrupamento dos anos 2000 a 2005, sendo o ano de 2006 um ano que não parece ter se agrupado por apresentar um valor muito menor de acidentes, destacando-se também um decréscimo no número de acidentes ano a ano.

Figura 5 – Apresentação gráfica das duas primeiras dimensões da ACP em relação ao número de acidentes por 10 mil veículos por ano



Fonte: Próprios autores (2020)

Foram determinadas a média e variância do número de acidentes por ano como na Tabela 4, bem como as matrizes de correlação da Tabela 5 e de variância-covariância da Tabela 6, a fim de facilitar a interpretação dos resultados e induzir a discussão sobre possíveis motivos para os agrupamentos.

A primeira observação feita com os alunos a partir dessas tabelas é que a diminuição da média de acidentes ao longo dos anos pode ser fruto de políticas públicas de redução de acidentes. Além disso, numericamente, verificou-se que o ano de 2006 apresentava a somatória total de acidentes e a média mais discrepantes, sendo também as menores. Tal fato se reflete no distanciamento gráfico desse ano dos demais e nas matrizes de correlação e variância-covariância apresentadas nas Tabelas 5 e 6.

Tabela 4 – Tabela exploratória dos dados de acidentes por ano

Ano	2000	2001	2002	2003	2004	2005	2006
Total	97,8	89,6	89,4	85,3	81,9	78,2	67,5
Média	3,15	2,89	2,88	2,75	2,64	2,52	2,18
Desvio Padrão	2,41	2,23	2,31	2,42	2,43	2,42	2,32

Fonte: Próprios autores (2020)

Tabela 5 – Matriz de correlação do número de acidentes por ano

	2000	2001	2002	2003	2004	2005	2006
2000	1,00	0,99	0,96	0,95	0,93	0,92	0,73
2001	0,99	1,00	0,98	0,98	0,97	0,96	0,82
2002	0,96	0,98	1,00	1,00	0,99	0,99	0,88
2003	0,95	0,98	1,00	1,00	0,99	0,99	0,89
2004	0,93	0,97	0,99	0,99	1,00	1,00	0,91
2005	0,92	0,96	0,99	0,99	1,00	1,00	0,92
2006	0,73	0,82	0,88	0,89	0,91	0,92	1,00

Fonte: Próprios autores (2020)

Tabela 6 – Matriz de variância-covariância do número de acidentes por ano

	2000	2001	2002	2003	2004	2005	2006
2000	5,82	5,31	5,34	5,53	5,45	5,34	4,06
2001	5,31	4,98	5,08	5,3	5,26	5,18	4,23
2002	5,34	5,08	5,35	5,57	5,59	5,53	4,71
2003	5,53	5,3	5,57	5,85	5,86	5,8	4,99
2004	5,45	5,26	5,59	5,86	5,93	5,88	5,15
2005	5,34	5,18	5,53	5,8	5,88	5,85	5,15
2006	4,06	4,23	4,71	4,99	5,15	5,15	5,37

Fonte: Próprios autores (2020)

Ressalta-se que foram apontados novos conhecimentos (a aplicação prática de conteúdos teóricos já vistos no Curso de Engenharia como Álgebra Linear, Cálculo Numérico e Estatística), entendidos como “aprender a saber” segundo Delors (1998) e, a execução de processamento computacional orientado, mesmo sem grande domínio da linguagem de programação particular, entendida como “aprender a fazer”, no sentido de Delors (1998).

Além disso, em relação aos resultados, os alunos entenderam e absorveram os conteúdos, que mais tarde foram retomados nas aulas de Processamento Digital de Imagens, onde se aborda novamente o conteúdo sobre ACP. Sendo também, posteriormente, afirmado por um deles que a experiência prática em sala adquirida por meio do estudo de caso aqui apresentado foi importante no desenvolvimento de atividades em seu trabalho como engenheiro civil.

Conclusões

Ficou evidenciada a empregabilidade e utilidade da ACP como ferramenta introdutória para análises multivariadas no ensino em cursos de engenharia. Pelo seu emprego, os estudantes puderam observar padrões e correlações existentes no exemplo dado em sala, sendo o mesmo o *ranking* de mortes no trânsito para diferentes países entre os anos 2000 e 2006. Além disso, o *software* livre escolhido, “RStudio”, permitiu aos alunos uma análise gráfica visual das mudanças ocorridas

ao longo dos anos, ficando evidente aos alunos os padrões de comportamento similares no período analisado dos grupos de países visualmente/graficamente observados, graças a sua facilidade de produção de gráficos. Por se tratar de um *software* livre os alunos puderam replicar sozinhos as metodologias aprendidas em sala sem necessidade de usar recursos financeiros próprios ou da universidade. Tais metodologias consolidaram os conhecimentos adquiridos e instigaram a busca por uma visão crítica dos dados observados.

Agradecimentos

Os autores agradecem ao Laboratório de Geoprocessamento do Departamento de Engenharia de Transportes da Escola Politécnica da Universidade de São Paulo, à Pró-Reitoria de Pós-Graduação pelo Programa de Aperfeiçoamento de Ensino (PAE), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - processo 88887.357063/2019-00 e a Charles Lincoln Kenji Yamamura, Doutorando do Programa de Pós-Graduação em Engenharia de Produção da EPUSP pelo suporte na língua inglesa.

Referências

ARA, A. B. **O ensino de Estatística e a busca do equilíbrio entre os aspectos determinísticos e aleatórios da realidade.** 2006. Tese (Doutorado em Educação) - Universidade de São Paulo, São Paulo, 2006.

ARANHA, M. L. A. **História da Educação**. 2. ed. São Paulo: Moderna, 1996.

BARTHOLOMEW, D. J. Principal Components Analysis. **International Encyclopedia of Education**, 3, p. 374-377, 2010. Disponível em: <https://doi.org/10.1016/B978-0-08-044894-7.01358-0>. Acesso em: 03 set. 2020.

DELORS, J (coord.). Os quatro pilares da educação. **Educação: um tesouro a descobrir**: Relatório para a UNESCO da Comissão Internacional sobre Educação para o Século XXI. Tradução de José Carlos Eufrázio. São Paulo: Cortez Editora. Brasília: Unesco, 1998.

DONY, R. 2001. Karhunen-loeve transform. *In*: **The transform and data compression handbook**, v. 1, 1-34. Boca Raton: CRC Press, 2001.

GAJEWSKI, B. J *et al*. Teaching Confirmatory Factor Analysis to Non-Statisticians: A Case Study for Estimating Composite Reliability of Psychometric Instruments. **Case Studies Bus Ind Gov Stat**, v. 5, n. 2, p. 88–101, 2014.

GOLD, P. A. **Traffic Saffety – Using Engineering to reduce acidentes**. Inter-American Development Bank, 1999.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 6, 417–441, 498–520, 1933.

JOLLIFFE, I. T. **Principal Component Analysis**. 2. ed. Berlin: Springer, 2002.

KLAŠNJA-MILIĆEVIĆ, A.; IVANOVIĆ, M.; BUDIMAC, Z. (2017). Data science in education: Big data and learning analytics. **Computer Applications In Engineering Education**, v. 25, p. 1066–1078, 2017. Disponível em: <https://doi.org/10.1002/cae.21844>. Acesso em: 03 set. 2020.

LINDNER, A. *et al*. Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study. **Geo-spatial Information Science**, v. 19, n. 4, p. 245-254, 2016. Disponível em: <https://doi.org/10.1080/10095020.2016.1260811>. Acesso em: 03 set. 2020.

LOPES, D. L. **Norma técnica 223 - Gestão da Informação e Redução de Acidentes de Trânsito no Brasil**, Companhia de Engenharia de Tráfego de São Paulo, 2012.

ORGANIZAÇÃO MUNDIAL DE SAÚDE (OMS). **Global status report on road safety**. Suíça, 2018.

PAPADIMITRIOU, E.; YANNIS, G. Is road safety management linked to road safety performance? **Accident Analysis and Prevention**, v. 59, p. 593–603, 2013. Disponível em: <https://doi.org/10.1016/j.aap.2013.07.015>. Acesso em: 03 set. 2020.

PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, v. 2, n. 6, p. 559–572, 1901.

PITOMBO, C. S.; GOMES, M. M. Study of Work-Travel Related Behavior Using Principal Component Analysis. **Open Journal of Statistics**, v. 4, p. 889-901, 2014. Disponível em: <https://doi.org/10.4236/ojs.2014.411084>. Acesso em: 03 set. 2020.

SIQUEIRA, A. D. O. *et al*. Estilos de Aprendizagem e Estratégias de Ensino em Engenharia. In: Congresso Brasileiro de Educação em Engenharia, 40., 2012 Belém, **Anais...**, Belém, 2012.

SOUZA, A. M. D.; POPPI, R. J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: um tutorial, parte I. **Química nova**, v. 35, n. 1, p. 223-229, 2012.

VALDERRAMA, L.; PAIVA, V. B.; MARCO, P. H.; VALDERRAMA, P. Proposta experimental didática para o ensino de análise de componentes principais. **Química nova**, v. 39, n .2, p .245-249, 2016.

WEGMAN, F. The future of road safety: A worldwide perspective. **International Association of Traffic and Safety Sciences**, n. 40, p. 66-71, 2016. Disponível em: <https://doi.org/10.1016/j.iatssr.2016.05.003>. Acesso em: 03 set. 2020.

WESTFALL, P. H.; ARIAS, A. L.; FULTON, L. V. Teaching principal components using correlations. **Multivariate Behavioral Research**, v. 52 n. 5, p. 648-660, 2017.

Recebido: 17/06/2020
Aprovado: 07/12/2020
Como citar: CARVALHO, F. S; MACHADO, C. A. S.; QUINTANILHA, J. A. Ensino de técnica de estatística multivariada para alunos de cursos de Engenharia. **Revista de Estudos e Pesquisas sobre Ensino Tecnológico (EDUCITEC)**, v. 6, e133420, 2020.