

# A UTILIZAÇÃO DA TÉCNICA PCA COMO FERRAMENTA DE OTIMIZAÇÃO DA CLASSIFICAÇÃO NÃO-SUPERVISIONADA

A. F. Carneiro<sup>1</sup>, T. F. Cordeiro<sup>2</sup>, I. S. Narvaes<sup>3</sup>

<sup>1</sup>Bolsista do Projeto MCTIAquaviário – CNPQ; Graduando do Curso de Engenharia Cartografia e de Agrimensura da Universidade Federal Rural da Amazônia – UFRA;

<sup>2</sup>Graduando do Curso de Engenharia Cartografia e de Agrimensura da Universidade Federal de Viçosa – UFV;

<sup>3</sup>Pesquisador do Instituto Nacional de Pesquisas Espaciais (INPE) no Centro Regional da Amazônia (CRA).

## Comissão IV

## RESUMO

O uso de técnicas de classificação automática tornam o mapeamento mais dinâmico a partir da possibilidade de serem identificadas feições, por meio de seus comportamentos espectral e estatístico, com relação a diferença existente no alvos imageados. Desta forma, neste trabalho é avaliado a acurácia da classificação não-supervisionada IsoDATA, utilizando como base um recorte de uma cena Landsat 8 – OLI, somente em reflectância e posteriormente aplicado a *PCA - Principal Component Analysis*. Verificou-se que, de maneira geral, a aplicação da *PCA* melhora o resultado da classificação, que foram avaliadas através dos índices de Exatidão Global e Kappa. Mesmo contendo algumas limitações, as classificações com o uso da *PCA* alcançaram um desempenho categorizado como bom, a tornando uma importante ferramenta no auxílio do mapeamento do uso e cobertura da Terra.

**Palavras chave:** Sensoriamento Remoto, Classificação não-supervisionada, ACP, IsoDATA.

## ABSTRACT

The use of automatic classification techniques makes the mapping more dynamic from the possibility of identifying features, through their spectral and statistical behavior, in relation to the existing difference in the target image. Based on that, this work evaluate the accuracy of the unsupervised IsoDATA classification, using a cut of a Landsat 8 - OLI scene, only in reflectance and subsequently applied a *PCA - Principal Component Analysis*. It was verified that, in a general way, the *PCA* application improves the classification result, which were evaluated through the Global Accuracy and Kappa indices. Even with some limitations, the classifications with the use of *PCA* achieved a performance categorized as good, making it an important tool in helping to map the use and coverage of the Earth.

**Keywords:** Remote Sensing, Unsupervised classification, *PCA*, IsoDATA.

## 1 - INTRODUÇÃO

A classificação automática de imagens é a forma em que o computador pode, de forma supervisionada ou não, mapear as feições existentes em uma cena de interesse com base em diversos parâmetros. No âmbito das Ciências Cartográficas, que abarcam as geotecnologias, é comum a utilização de técnicas de mapeamento, cada vez mais automatizadas, a fim de propiciar produtos de forma mais ágil e com precisão compatível com os principais programas de monitoramento em larga escala.

Com o avanço da tecnologia e o surgimento de computadores cada vez mais robustos, o custo

computacional para a interpretação dos dados tende a diminuir numa escala inversa ou diretamente proporcional a diversos aspectos, como a capacidade do computador, quantidade de feições a serem mapeadas, número de iterações, tamanho da cena, resolução espacial e processos realizados a fim de otimizar essa tarefa de identificar as feições.

Para programas de monitoramento, que visam avaliar as alterações na cobertura florestal da Amazônia, à exemplo do DETER (Diniz et al, 2015), PRODES e DEGRAD ou o padrão de uso e cobertura em áreas anteriormente desmatadas, denominado de TerraClass (Almeida et al., 2016) este tipo de avaliação é fundamental, afim de utilizar a informação que se

pode extrair de um produto originado de um sistema sensor e usá-la para classificar feições características das classes, que se torna possível a partir da leitura da informação contida no conjunto de pixels que formam a feição.

Neste contexto, este trabalho visa a utilização da PCA - *Principal Component Analysis* - de forma prévia nas imagens a serem classificadas pela técnica de classificação não-supervisionada IsoDATA, bem como sem a aplicação da PCA, afim de avaliar a acurácia das classificações em uma região com intensa dinâmica de mudança de uso e cobertura da terra na Amazônia e fornecer subsídios para a melhoria dos programas de monitoramento da Amazônia, em especial o projeto TerraClass.

## 2 - FUNDAMENTAÇÃO TEÓRICA

O processo de classificação é baseado na capacidade do computador de extrair informações para reconhecer padrões, presentes em imagens, utilizados para mapear áreas da superfície terrestre (Freitas e Pancher, 2011). Apesar da existência de diversos processos de classificação, as formas mais difundidas entre os usuários de sensoriamento remoto, adotada pelos softwares comerciais de processamento de imagens estão divididos entre os métodos supervisionados e não-supervisionados. (Meneses e Almeida, 2012).

Para a classificação supervisionada é necessário que haja conhecimentos prévios das classes dos alvos, com a finalidade de classificar a imagem nas classes-alvo do analista. O algoritmo necessita ser treinado para poder distinguir as classes uma das outras, onde neste método, é necessário colher amostras das classes que se deseja obter a partir do processamento. Assim, há vários métodos para classificação supervisionada como paralelepípedo, distância mínima, distância de Mahalanobis e máxima verossimilhança (Meneses e Almeida, 2012).

Já as técnicas de classificação automática, não-supervisionada, baseiam-se no princípio de que o algoritmo computacional é capaz de identificar por si só as classes dentro de um conjunto de dados. Esse tipo de classificação é frequentemente realizada por meio de métodos de agrupamentos (*clustering*). No entanto, tais técnicas possuem limitações, por serem utilizados métodos de agrupamentos particionais, onde as funções objetivo usadas por tais métodos partem do pressuposto de que o número de agrupamentos ou classes é conhecido a priori. Sendo assim, no caso de ser inserido um valor não correspondente ao número real de classes, o método irá impor, pelo uso de técnicas de otimização, o valor escolhido, para o agrupamento de dados (Gonçalves et al., 2008), o que dependendo da acurácia alcançada pode ser um balizador para a obtenção de produtos de forma mais rápida em programas de monitoramento em larga escala, logicamente respeitando uma precisão aceitável.

A maioria dos programas de geoprocessamento e de sensoriamento remoto possuem os métodos não-supervisionados, tais como o K-médias e ISODATA (Gonçalves et al., 2008).

O método *clustering* denominado de ISODATA usa a distância espectral num processo feito em três etapas: 1 – classifica os pixels de forma iterativa; 2 – após o primeiro, redefine os critérios para cada classe; 3 – reclassifica a imagem até os padrões de distância espectral dos dados apareçam e fundamenta-se no princípio em que todos os agrupamentos formados na imagem devem ser compactos, com todos os pixels agrupados em volta da média. As condições para a divisão e fusão de agrupamentos giram em torno do desvio padrão da resposta em relação a média e em relação a distância mínima que existe entre os centros dos agrupamentos (Meneses e Almeida, 2012).

No estudo de uma forma para melhorar os resultados dos classificadores, a análise de componentes principais (ACP) é uma técnica estatística que busca tornar um conjunto de imagens, linearmente não correlacionado. Ao rotacioná-los em um espaço de atributos, através de uma transformação ortogonal, são convertidos um conjunto de observação de variáveis possivelmente correlacionadas num conjunto de valores de variáveis linearmente não correlacionadas, as quais explicam grande parte da variância dos dados, em componentes principais (Neto e Moita, 1997). Cada componente principal é uma combinação linear de todas as variáveis originais, 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, sendo que o tamanho do conjunto de imagens de entrada é o mesmo tamanho do conjunto de saída, sendo que retém em ordem de estimação a maior quantidade de informações possível com o mínimo de perda (Varela, 2008).

Esta técnica deposita em ordem de componentes principais as informações, portanto a chamada primeira componente principal ou (CP1) contém as informações espectrais mais significativas comuns entre as imagens de entrada, a segunda os dados um pouco menos significativas, e assim até a última componente que apresenta o menor percentual de informação para o conjunto de dados (Crosta, 1992). Desta forma é possível que, com o conhecimento do tipo de distribuição de níveis de cinza e seu desvio padrão na direção ortogonal, seja possível verificar os autovalores da matriz de correlação e assim melhorar os resultados da classificação.

## 3 – MATERIAIS E MÉTODOS

O trabalho foi realizado a partir de um recorte de uma cena Landsat-8 (Sensor: OLI; Órbita/Ponto: 227/062; Data: 30/10/2014) que abrange os municípios de Belterra e Mojuí dos Campos, localizados no Estado do Pará. A cena foi disponibilizada, de forma gratuita, no sítio da USGS – *United States Geological Survey* – pela plataforma *EarthExplorer*, que possibilita aos

usuários o *download* de imagens já corrigidas atmosféricamente através do método ToA (reflectância no topo da atmosfera), disponível em <<https://earthexplorer.usgs.gov/>>.

A área de interesse foi estabelecida pelo recorte da cena, compreendido entre as coordenadas 55° 5' 0.210" W 2° 40' 22.785" S; 54° 35' 19.232" W 3° 0' 53.608" S (Figura 1).

Foram utilizadas as bandas de 2 à 7, que corresponde ao comprimento de onda do visível ao infravermelho médio em ordem crescente. Em seguida, após o empilhamento das bandas (*layer stack*).

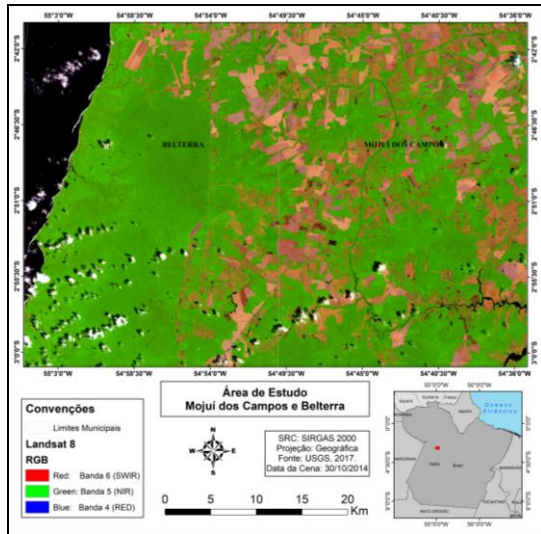


Figura 1 - Mapa de Localização.

Posteriormente, foi aplicado sobre este produto da técnica de ACP, onde foram geradas as componentes em ordem de correlação, sendo que posteriormente foram utilizadas as componentes 1, 2, 3 e 4, individualmente e também agrupadas como bandas (Figura 2), para a aplicação da classificação ISODATA, a ser comparado com o mesmo método de classificação nas bandas espectrais empilhadas sem a aplicação da ACP.

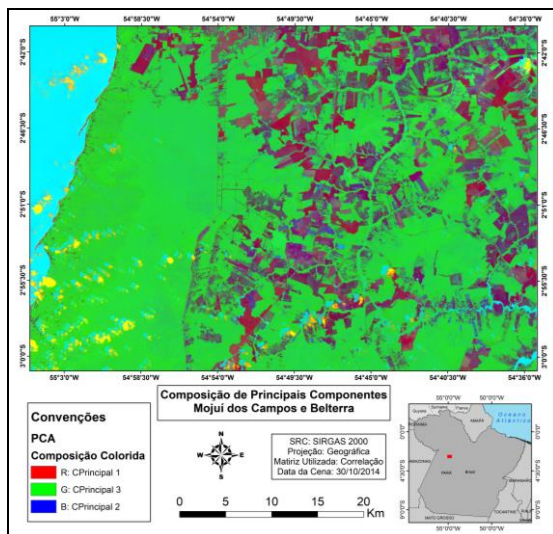


Figura 2 – Aplicação da técnica de ACP. Composição R (ACP1); G (ACP3) e; B (ACP2).

Como forma de validar o resultado e instruir os classificadores sobre a quantidade de classes a serem utilizadas nesta análise para a área em questão, utilizou-se como base a classificação fornecida pelo projeto TerraClass (Almeida et al., 2016), cujo objetivo é realizar o monitoramento dos padrões de uso e ocupação do solo de áreas anteriormente desmatadas e detectadas no PRODES, por meio de interpretação via fotointérprete de produtos oriundos do sensoriamento remoto, utilizadas como verdade de campo para a classificação automática.

As classes do monitoramento de uso e cobertura da Amazônia – TerraClass (Almeida et al., 2016), totalizam 12 classes encontradas na área de estudo: Agricultura Anual; Área não Observada; Área Urbana; Desflorestamento\_2014; Floresta; Hidrografia; Mosaico de Ocupações; Outros; Pasto Limpo; Pasto Sujo; Regeneração com Pasto e Vegetação Secundária.

Na composição do TerraClass, algumas classes temáticas foram importadas do banco de dados do programa PRODES Floresta; Não Floresta; Desmatamento (do ano) e Hidrografia (Câmara et al., 2006). Alguns autores como Duarte et al. (2002) e De Espindola et al. (2004) afirmam acerca do deslocamento geométrico que existe nos dados oriundos do projeto PRODES, o que pode influenciar no resultado as classificações na região de contado com a máscara destas camadas temáticas oriunda do PRODES estabelecida nesta análise.

Em adição, também para os dados do projeto TerraClass foi aplicada uma máscara para a classe temática de “mosaico de ocupações”, por esta ser representada por uma associação de diversas modalidades de uso da terra e que, devido à escala de trabalho adotada, não ser possível executar a discriminação dos diferentes componentes da paisagem, tais como a agricultura tradicional e de pastagem para criação extensiva de gado (Coutinho et al., 2013), não sendo possível discerní-la por meio do classificador não-supervisionado aplicado.

O mesmo se aplicou para a classe de “Área Urbana” pois a classe que compõe a paisagem não se apresenta densidade alta de construções, ou seja, com alternância com áreas de vegetação e solo exposto, o que acarreta a formação de distribuição de níveis de cinza muito distintos, o que dificulta a sua correta discriminação.

Por conta do comportamento espectral similar entre as classes de Pasto Sujo, Regeneração com Pasto e Vegetação Secundária estas foram consideradas áreas de Vegetação Secundária, para o classificador, tendo sido agrupadas na classe vegetação secundária.

Após o agrupamento das classes, estas foram rotuladas às classes temáticas de Agricultura Anual; Área não observada; Pasto Limpo e Vegetação Secundária.

Para o cálculo do desempenho das diferentes técnicas aplicadas ao classificador gerado, foi necessário agrupar classes de referência afim de compatibilizá-las a classificação geradas, obtendo desta forma mesmo número de classes entre ambas, tendo em vista que o classificador automático não se mostrou eficaz na discriminação das classes de Pasto Sujo, Regeneração com Pasto e Outros.

A partir do resultado da ACP e do agrupamento das bandas originais, as imagens foram classificadas a fim de comparar a eficácia e acurácia dos classificadores. Os índices utilizados para a avaliação das classificações foram o Índice Kappa, Índice de Exatidão Global, além da geração de uma matriz de Confusão (Figueiredo e Vieira, 2007).

O desempenho da classificação para o valor *Kappa* obtido, os quais variam de 0 a 1, aceitos na comunidade científica variam entre classificações com desempenho de péssimo a excelente.

#### 4 – RESULTADOS

Os diferentes métodos e técnicas utilizadas produziram um número de classes diferentes, decorrente das variações na radiometria dos produtos gerados, agrupados para compor o mesmo número de classes de referência (04 classes). Assim, o classificador IsoDATA foi configurado com um número mínimo de classes igual a referência e máximo de classes a serem mapeadas, estabelecidas em 100, levando em conta as iterações produzidas, inferindo um número máximo de classes, variáveis de acordo com a radiometria da imagem, número mínimo de pixels por classe, média e desvio padrão dos níveis de cinza, resultando em número máximo de classes variáveis de acordo com a técnica aplicada ao classificador (Figura 3).

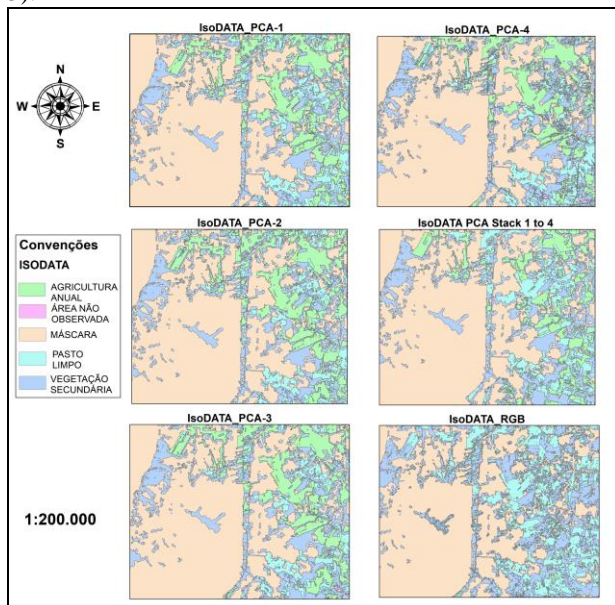


Figura 3 - Resultados das Classificações IsoDATA.

As classificações produziram resultados em geral elencadas com desempenho considerado bom, de

acordo com Fonseca (2000). Apenas a classificação utilizando apenas a segunda componente principal obteve valor de kappa considerado com desempenho razoável (Tabela 1).

Tabela 1 - Índices de acurácia global e índice kappa, para as classificações geradas.

Classificação	Acurácia Global (%)	Kappa
PCA_1to4	72,7675	0,5848
RGB	72,7361	0,5847
PCA_1	70,87	0,5551
PCA_4	65,6846	0,4694
PCA_3	62,9585	0,4314
PCA_2	54,6886	0,2881

Dessa forma, a melhor classificação produziu uma matriz de confusão capaz de mostrar as características deste resultado (Tabela 2).

Tabela 2 - Matriz de confusão em pixels e percentual para a classificação IsoData PCA 1 a 4; A.A = Agricultura Anual; A.n.O = área não observada; P. L. = Pasto Limpo; e VS = vegetação secundária. Valores percentuais entre parênteses.

Matriz de Confusão					
TerraClass Adaptado x PCA Empilhada (1to4)					
Class	A.A	A.n.O	P.L	V.S	Total
A.A	<b>194991</b> (69,23%)	845 (5,71%)	41207 (12,49%)	12920 (2,79%)	249963 (22,94%)
A.n.O	1150 (0,41%)	<b>3928</b> (26,54%)	1715 (0,52%)	2028 (0,44%)	8821 (0,81%)
P.L	41982 (14,91%)	6247 (42,21%)	<b>224187</b> (67,93%)	78454 (16,93%)	350870 (32,19%)
V.S	43529 (15,45%)	3781 (25,55%)	62940 (19,07%)	<b>369961</b> (79,84%)	480211 (44,06%)
Total	281652 (100,00%)	14801 (100,00%)	330049 (100,00%)	463363 (100,00%)	<b>1089865</b> (100,00%)

Embora o desempenho das classificações tenham sido similares, a aplicação da análise de componentes principais foi responsável por valores maiores de acurácia e desempenho da classificação não-supervisionada. Os trabalhos de Freitas e Cruz (2005) também elucidam a melhora na qualidade das classificações quando utilizada a técnica de análise de componentes principais. Souza et al (2009) complementa ainda, que tal técnica é um bom método para avaliar mudanças na cobertura da terra.

Mesmo que todas as classificações tenham sido consideradas satisfatórias, um valor considerável de áreas de agricultura anual foram erroneamente classificadas como pasto limpo, onde há erro de inclusão (Erro do tipo 1) atrelado as características radiométricas semelhantes entre estes dois tipos de uso, ao passo que um valor considerável de áreas que deveriam ser de agricultura anual. Uma parte considerável das áreas de vegetação secundária, ou seja, as áreas classificadas como pasto limpo e agricultura anual não foram classificadas corretamente como pertencendo a esta classe. Já as áreas de pasto limpo obtiveram erros de inclusão significativo principalmente como áreas de vegetação secundária.

Os maiores percentuais de confusão das classes foram: falso positivo de Vegetação Secundária (15,45%) para a classe de Agricultura Anual; falso positivo de Pasto Limpo (42,21%) para a classe de Área não Observada; falso positivo de Vegetação Secundária (19,07%) para a classe de Pasto Limpo; e falso positivo de Pasto Limpo (16,93%) para a classe de Vegetação Secundária,

A partir desta análise, os resultados que obtiveram o maior percentual de acerto, em ordem decrescente, foram: Vegetação Secundária (79,84%), Agricultura Anual (69,23%), Pasto Limpo (67,93%) e Área não Observada (26,54%),

#### 4 – CONCLUSÕES

De acordo com os resultados gerados é possível concluir que:

Elementos que possuem pixels de diversas características radiométricas seriam melhor classificados, na aplicação de uma técnica de classificação supervisionada;

A exceção das bandas do termal e pancromática a utilização das demais, com a aplicação da técnica da ACP, foi possível obter uma boa acurácia do mapeamento do uso e cobertura do solo para imagens de média resolução espacial.

Recomenda-se a continuidade do estudo, com a aplicação de classificadores supervisionados, o qual levam em consideração as amostras de treinamento definidas pelo intérprete, mais precisos do ponto de vista estatístico.

#### REFERÊNCIAS BIBLIOGRÁFICAS

Almeida, C. A. et al. 2016. High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data, ACTA AMAZONICA, v, 46, p, 291-302, 2016,

CÂMARA, Gilberto; et al. 2006 Metodologia para o cálculo da taxa anual de desmatamento na Amazônia Legal, São José dos Campos: INPE, 2006.

Coutinho et al. 2013. Uso e cobertura da terra nas áreas desflorestadas da Amazônia Legal: TerraClass 2008, Brasília, DF: Embrapa; Belém: INPE, 2013, 108p.: il, Color, ISBN 978-85-7035-180-7.

Crosta, A. P. 1992. Processamento digital de imagens de sensoriamento remoto, Unicamp/Instituto de Geociencias, São Paulo, 169 p.

DE ESPINDOLA, Giovana Mira; et al. 2004. Aplicação da metodologia do PRODES digital em imagens CCD/CBERS-2, Simpósio Brasileiro de Sensoriamento Remoto, 2004.

DUARTE, Valdete et al. 2002. Metodologia para corrigir as distorções observadas entre os mapeamentos temporais do projeto PRODES Digital, São José dos Campos: INPE, 2002.

Figueiredo, G. C.; Vieira, C. A. O, Estudo do comportamento dos índices de Exatidão Global, Kappa e Tau, comumente usados para avaliar a classificação de imagens do sensoriamento remoto, Simpósio Brasileiro de Sensoriamento Remoto, v, 13, p, 5755-5762, 2007.

FONSECA, L. M. G. Processamento digital de imagens. Instituto Nacional de Pesquisas Espaciais (INPE), 2000. 105p.

Freitas, M. I. C, de; Pancher, A. M., 2011, Classificação de imagens, Unesp, Rio Claro, 33 slides, color, Disponível em: <[http://www.rc.unesp.br/igce/planejamento/download/isa-bel/sist\\_inf\\_geografica/Aula11/classificacao\\_imagens.pdf](http://www.rc.unesp.br/igce/planejamento/download/isa-bel/sist_inf_geografica/Aula11/classificacao_imagens.pdf)>, Acesso em: 21 jul, 2017.

Freitas, S. R., & Cruz, C. B. M. 2005, Análise de Componentes Principais e Modelo Linear de Mistura na discriminação de classes de vegetação na Mata Atlântica. Anais XII Simpósio Brasileiro de Sensoriamento Remoto, 1529-1536.

Gonçalves, M. L. et al., 2008. Classificação não-supervisionada de imagens de sensores remotos utilizando redes neurais auto-organizáveis e métodos de agrupamentos hierárquicos, Revista Brasileira de Cartografia. v. 1, n. 60.

Meneses, P. R.; Almeida, T, de., 2012. Introdução ao processamento de imagens de sensoriamento remoto, Embrapa Cerrados - Livros técnicos (INFOTECA-E),

NETO, J. M.; MOITA, G. C. Uma introdução à análise exploratória de dados multivariados, Química nova, v, 21, n, 4, p, 467-469, 1998.

Souza, S. F. et al. 2009. Utilização de análise por componentes principais (ACP) no diagnóstico das alterações da cobertura vegetal densa na bacia do rio Natuba - PE. In: Simpósio Brasileiro De Sensoriamento Remoto, XIV, Natal. Anais XIV SBSR. [s.i]: Inpe, 2009. p. 7189 - 7196.

Varella, C. A. A. 2008. Análise de Componentes Principais, Disponível em: <[http://www.ufrrj.br/institutos/it/deng/varella/Downloads/multivariada\\_aplicada\\_as\\_ciencias\\_agrarias/Aulas/analise\\_de\\_componentes\\_principais.pdf](http://www.ufrrj.br/institutos/it/deng/varella/Downloads/multivariada_aplicada_as_ciencias_agrarias/Aulas/analise_de_componentes_principais.pdf)>, Acesso em: 21 jul, 2017.