Aplicação de análise exploratória multivariada em dados complexos: Revisão Sistemática

Elizabete Yanase Hirabara[a,*], Nadia Puchalski Kozievitch[a], Adolfo Gustavo Serra Seca Neto[a]

[a] *Universidade Tecnológica Federal Do Paraná*
Departamento Acadêmico de Informática
Avenida Sete de Setembro, 3165, 80.230-901 – Curitiba – PR – Brazil

Abstract

Contexto: Com o aumento da geração de informação nos dias atuais, cresce proporcionalmente a dificuldade de analisar e entender o grau de relacionamento entre cada dado a fim de gerar e extrair conhecimento. Este problema é ainda maior quando se tratam de dados oriundos das mais diversas fontes tecnológicas ou formatos, sendo em ambos casos conhecidos como dados complexos. Para tal, é essencial a aplicação de métodos de análise exploratória multivariada, sendo os mais utilizados a Análise de Agrupamento Hierárquico ou *hierarchical cluster analysis* (HCA) e o Agrupamento de Componentes *Principais ou Principal Component Analysis* (PCA), sendo métodos que identificam as variáveis que agregam conhecimento em relação com as demais e verificam se esta relação é forte ou não em comparação com outras variáveis encontradas.

Método: Neste estudo, foi realizado uma revisão sistemática, considerando trabalhos publicados entre o período de 2010 a 2016.

Resultados: Nos estudos analisados cerca de 70,2% aplicaram o PCA, destes, aproximadamente 57,7% são da área de Ciências Exatas e da Terra, sendo que 93,3% utilizaram experimentos para validar a solução.

Conclusão: Com os resultados encontrados, o método de PCA é o mais

*Corresponding author
Email addresses: ehirabara@gmail.com (Elizabete Yanase Hirabara), nadiap@utfpr.edu.br (Nadia Puchalski Kozievitch), adolfo@utfpr.edu.br (Adolfo Gustavo Serra Seca Neto)

aplicado atualmente, sua utilização tem crescido desde 2014 em muitas áreas de conhecimento, mas este método mostrou ser ineficaz quando o conjunto de dados é muito grande, podendo tornar-se impossível de processar ou gerar padrões cíclicos.
Keywords: Análise Multivariada, Dados Complexos, Revisão sistemática.

1. Introdução

De acordo com um dos Grandes Desafios da Computação no Brasil: 2006 - 2016, lançado pela Sociedade Brasileira de Computação (SBC), o grande volume de dados gerados diariamente tem aumentado, tais dados que são disponibilizados em formatos diversos dificultam a gestão de informação a fim de gerar conhecimento.

A dificuldade em processar tais informações ocorre muitas vezes por falta de recursos tecnológicos, financeiros e também em decorrência da dificuldade em integrar dados oriundos das mais diversas fontes tecnológicas ou formatos, sendo em ambos casos definidos como dados complexos [1].

Pelo grande volume dos dados a analisar e entender o grau de relacionamento entre as variáveis, a aplicação de técnicas estatísticas de forma manual em todos os casos pode ser humanamente impossível. Sendo necessário alguns métodos, como por exemplo, a Análise exploratória multivariada, a qual fornece uma ideia de como variáveis relacionam-se e distribuem-se, uma ferramenta aplicada para as diferentes áreas do conhecimento, utilizada para medir, explicar e prever o grau de relação entre estas variáveis, possibilitando aplicar em uma única análise aquilo em que deveria ser aplicado múltiplas análises univariadas [2].

Um exemplo seria a aplicação desta técnica para validar dados de vários documentos, de diferentes formatos, com a finalidade de verificar possíveis incidência de irregularidades e detectar padrões [3].

Dentre os principais métodos para esta averiguação, é possível destacar a Análise de agrupamento Hierárquico (HCA) ou *Principais ou Principal Component Analysis* e Análise de componentes principais ou *Principais ou Principal*

Component Analysis (PCA), por sua aplicação multidisciplinar [4], que serão relatadas mais profundamente na seção 2.

Desse modo, é essencial ter conhecimento teórico dos diferentes métodos aplicados atualmente para a coleta, análise de dados e detecção de padrões destes. Este trabalho selecionou e classificou os estudos atuais realizados em diversas áreas de pesquisa aplicando a análise exploratória multivariada, agrupando estas abordagens de acordo com os métodos de HCA e PCA(i), (ii) áreas de pesquisa, e (iii) abordagens atualmente mais utilizadas nas áreas avaliadas. Para tal propósito, foram considerados os trabalhos publicados entre 2010 e 2016.

O processo de revisão sistemática utilizado neste estudo identificou primariamente 138 trabalhos; 37 deles foram selecionados e discutidos em detalhe. Com a análise detalhada de cada um dos trabalhos, esta revisão fornece ênfase nos formatos de dados utilizados e a área de pesquisa de aplicação.

Depois de consolidados, os resultados desta revisão mostrou que a maior parte dos trabalhos selecionados utilizam método de análise multivariada PCA e estes foram em sua maioria publicados no ano de 2013.

As seções deste trabalho foram organizadas da seguinte forma: Seção 2 contém uma definição detalhadas sobre os dados complexos, análise de correlação multivariada e principais técnicas que a utilizam. Seção 3 apresenta o método utilizado para o desenvolvimento deste trabalho. Seção 4 disponibiliza os resultados das análises. A Seção 5 apresenta a discussão. Finalmente, na última seção, a conclusão.

2. Contexto

2.1. Dados Complexos

A informação tem sido a principal ferramenta utilizada pelas empresas e pelas organizações institucionais para a descoberta e divulgação de conhecimento. Para que isto seja possível, tais dados, gerados em volumes expressivos devem ser armazenados, adaptados, analisados e muitas vezes compreendidos em diferentes

padrões, para que seja possível um apoio à tomada de decisão [5]. Tal processo, muitas vezes têm como fonte inicial dados oriundos de bases com tecnologias, padrões e fins adversos, sem qualquer forma de padronização. Quando integrados, tais informações são definidas como dados complexos [1].

2.2. Análise Exploratória de Dados

Análise Exploratória de Dados é estudo profundo dos dados a serem estudados, validando, resumindo, organizando, analisando, interpretando e extraindo conhecimento a partir destes. Sendo que esta análise pode ser aplicada a uma variável ou multiariáveis, sendo dividida entre análise univariada e análise multivariada [4].

2.2.1. Análise Univariada

Um método de análise descritiva ou dedutiva a qual analisa uma só variável separadamente. Tendo como objetivo resumir as principais características em um grupo de dados fazendo uso de tabelas, gráficos e resumos numéricos [3].

2.2.2. Análise Multivariada

Pelo grande volume dos dados, sendo a heurística difícil de detectar, é necessário a aplicação de uma análise para medir a relação entre as variáveis já conhecidas, podendo esta ser fraca ou forte e medida através de um coeficiente de correlação, conhecida matematicamente como análise de correlação multivariada. Esta análise pode ser aplicada tanto para prever possíveis comportamentos como detectar anomalias em padrões pré-existentes [6].

2.2.3. Análise de agrupamento Hierárquico (HCA)

A análise de agrupamento hierárquico ou HCA como é conhecida, é uma análise exploratória multivariada que define-se como o tratamento matemático de uma amostra disposta como um ponto no espaço multidimensional de acordo com as variáveis escolhidas, interligando as variáveis por suas associações [3].

Estas associações ou relações são calculadas através da distância entre as variáveis, como é possível ser visualizada na equação 1. Podendo ser definida

como a distância de Xi, em relação a A, onde "Xi" é um vetor aleatório de um determinado espaço vetorial e "A" é um ponto qualquer, é igual à soma da diferença de Xi e A ao quadrado, onde "i" é um número entre 1 e N, e este por sua vez é um número real, não negativo [4].

$$d(Xi.A) = \sqrt{\sum_{i}^{N}(Xi - A)^2} \tag{1}$$

2.2.4. Análise de componentes principais (PCA)

A análise de componentes principais ou PCA, é a análise dos dados com a finalidade de reduzir, eliminar sobreposições e escolher formas representativas de dados a partir da combinação linear das variáveis de origem [7].

3. Método

Esta revisão sistemática foi realizada seguindo um guideline proposto por [8], com o intuito de evoluir o conhecimento sobre o tema explorado e responder às questões de pesquisa apresentadas na subseção 3.2.

3.1. Questões de pesquisa

As questões a serem respondidas por esta revisão estão apresentadas a seguir.

PQ1 - Qual dos métodos estudados é o mais aplicado para análise de informações em dados complexos?

PQ2 - Qual a área de conhecimento que mais utiliza a análise multivariada?

Dentre os pontos fracos identificados para o desenvolvimento do estudo, evidencia-se a aplicação destas soluções teóricas em problemas reais do cotidiano. Portanto a questão a respeito da limitações está disponibilizada a seguir.

LQ1 - Os trabalhos revisados eram validados através de experiências?

3.2. Critérios de Inclusão e Exclusão

Após o processo de pesquisa, com o resultado obtido foi aplicado como critério de seleção trabalhos que apresentaram um dos métodos estudados (HCA

ou PCA) e dados complexos, considerando a definição apresentada na seção 2, sendo ambos (método e dados complexos) encontrados na parte de resumo ou título. Para a aplicação dos critérios de inclusão o título e o resumo do artigo foram considerados. Para os critérios de exclusão, foram excluídos os trabalhos que tiveram sua publicação cancelada, duplicados ou que não foi possível obter acesso ao trabalho completo.

3.3. Processo de Pesquisa

Para o processo de revisão, foram selecionados trabalhos publicados em conferências e periódicos entre os anos de 2010 a 2016, disponibilizadas nas bases de artigos científicos, os escolhidos podem ser visualizados na tabela 1.

Table 1: Bases de pesquisa escolhidas

Base escolhida	Data de pesquisa
IEEE Xplore	03/08/2016
ACM Digital Library	03/08/2016
ScienceDirect	03/08/2016

Durante o processo de revisão sistemática utilizados neste estudo foram inicialmente identificados 138 artigos, conforme indicado na Figura 1, 37 deles foram selecionados e discutidos em detalhe após a aplicação dos critérios de seleção. Foram utilizados, para busca dos artigos, os seguintes descritores e suas combinações nas línguas portuguesa e inglesa: "Análises de Componente principais" AND "Dados complexos", "Agrupamento hierárquico" AND "Dados complexos".

3.4. Informações extraídas

As informações extraídas e analisadas de cada trabalho podem ser visualizadas a seguir.

- Título, autor, ano e local de publicação;

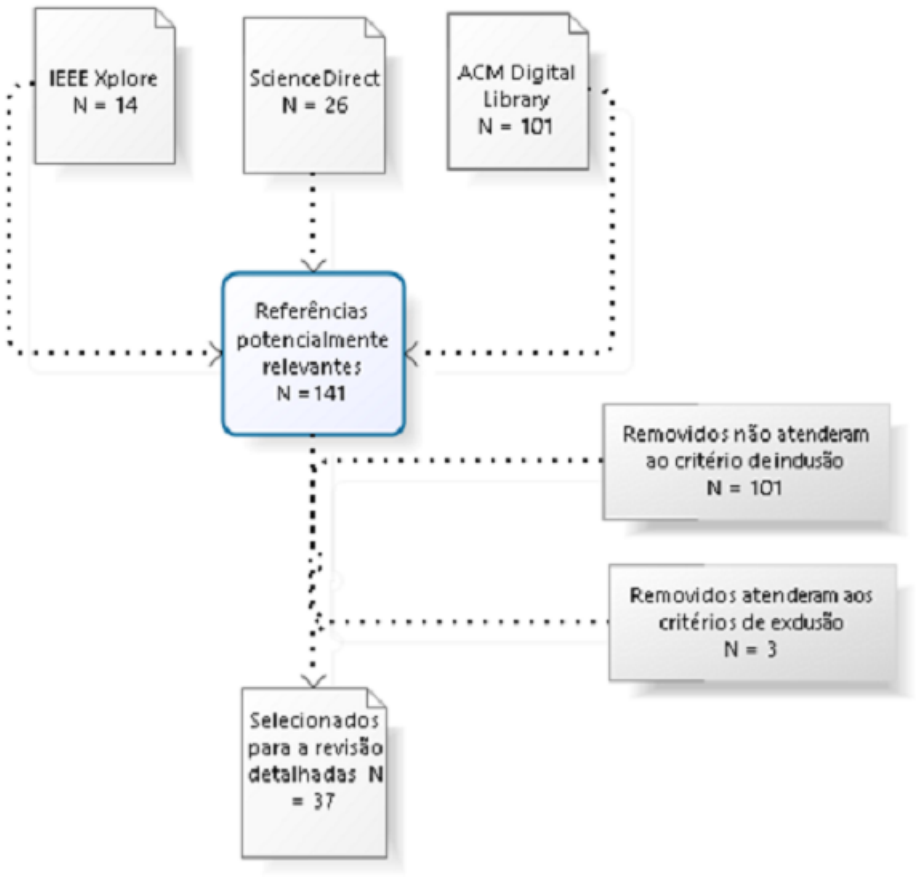

Figure 1: Processo de revisão sistemática

- Detalhes sobre o problema abordado;
- Limitações;
- Tipo do método aplicado (HCA ou PCA);
- Área de pesquisa; e
- Existência de experimentos;

4. Resultado da análise

Um total de 37 trabalhos foram selecionados para esta revisão sistemática, sendo estes publicados entre 2010 e 2016, todos abordando a aplicação de métodos da análise explorativa multivariada HCA ou PCA com dados complexos, compostos por 25 estudos utilizando métodos PCA, 11 utilizando HCA e 1 com ambos, como é possível visualizar na Figura 2

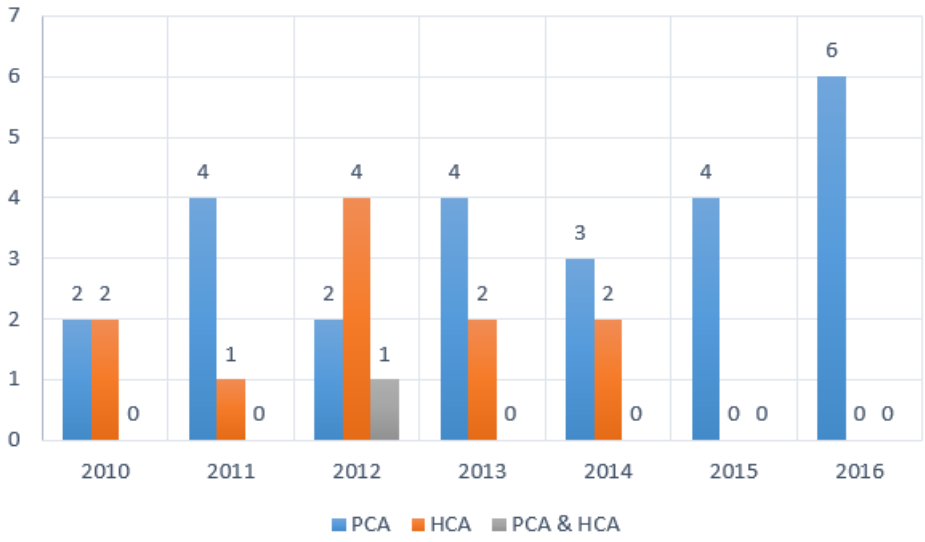

Figure 2: Quantidade de publicações entre os anos de 2010 a 2016 separadas por métodos (HCA e PCA)

Na Tabela 2 estão listados com mais detalhes os trabalhos que foram analisados nesta Revisão Sistemática, suas áreas de pesquisa ou conhecimento conforme definições e distribuições do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) [1] e se em seus trabalhos foram implementados experimentos.

[1] http://www.cnpq.br/documents/10157/186158/TabeladeAreasdoConhecimento.pdf - último acesso em 17/08/2016

Table 2: Lista de trabalhos revisados com detalhes.

Artigo	**Área de pesquisa/conhecimento**	**Utilizou experiências?**
[9]	Ciências da Saúde	Sim
[10]	Ciências Exatas e da Terra	Sim
[11]	Ciências Exatas e da Terra	Sim
[12]	Ciências Exatas e da Terra	Sim
[13]	Ciências Exatas e da Terra	Sim
[14]	Ciências Exatas e da Terra	Sim
[15]	Ciências Exatas e da Terra	Sim
[16]	Ciências Exatas e da Terra	Sim
[17]	Ciências Exatas e da Terra	Sim
[18]	Ciências Exatas e da Terra	Sim
[19]	Ciências Humanas	Não
[20]	Ciências Biológicas	Sim
[21]	Ciências Exatas e da Terra	Sim
[22]	Engenharias	Não
[23]	Ciências Biológicas	Sim
[24]	Ciências Biológicas	Sim
[25]	Ciências Exatas e da Terra	Sim
[26]	Ciências Agrárias	Sim
[27]	Ciências Exatas e da Terra	Não
[28]	Ciências Biológicas	Sim
[29]	Ciências Exatas e da Terra	Sim
[22]	Ciências Exatas e da Terra	Sim
[30]	Ciências Agrárias	Não
[31]	Ciências da Saúde	Sim
[32]	Ciências da Saúde	Sim
[33]	Ciências Exatas e da Terra	Sim
[34]	Ciências Exatas e da Terra	Sim
[35]	Ciências Sociais Aplicadas	Não

Artigo	Área de pesquisa/conhecimento	Utilizou experiências?
[36]	Engenharias	Não
[37]	Ciências Exatas e da Terra	Sim
[38]	Ciências da Saúde	Sim
[39]	Ciências Exatas e da Terra	Sim
[40]	Ciências Exatas e da Terra	Sim
[41]	Ciências Exatas e da Terra	Não
[42]	Ciências Agrárias	Sim
[43]	Ciências Exatas e da Terra	Sim
[44]	Ciências Exatas e da Terra	Sim

Com mais detalhes, separando por métodos, na Figura 3 e Figura 4, é possível verificar as quantidade que foram publicadas por ano em cada área de conhecimento.

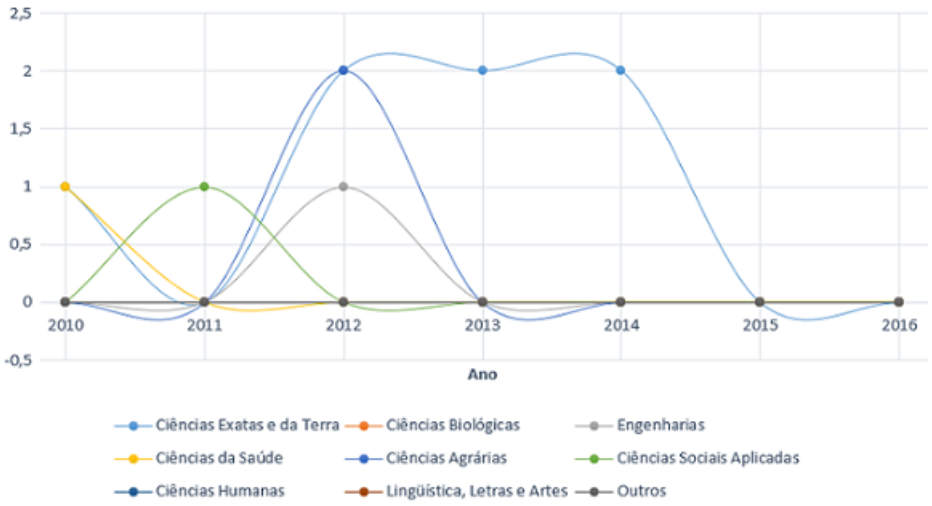

Figure 3: Publicações de áreas de conhecimento entre os anos de 2010 a 2016 (HCA)

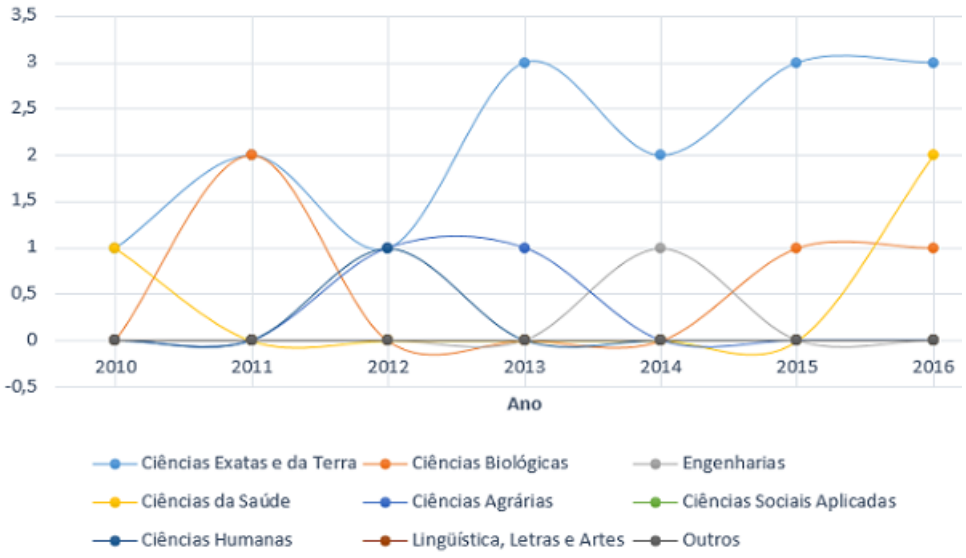

Figure 4: Publicações de áreas de conhecimento entre os anos de 2010 a 2016 (PCA)

Foi ainda verificado que em cerca de 84,6% dos trabalhos com método PCA e 66,7% dos trabalhos com método HCA, a validação dos resultados foram feitos através de experimentos, este resultado pode ser visualizado nas Figuras 5 e 6 nesta ordem.

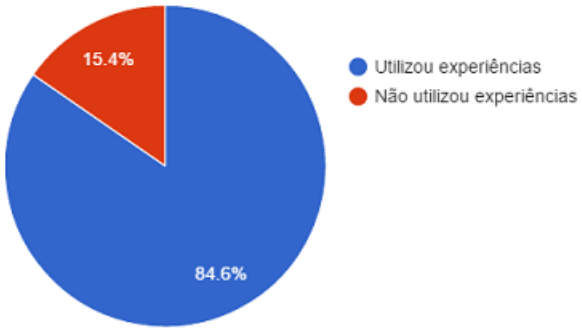

Figure 5: Gráfico de publicações que aplicaram método PCA por validação através de experiências

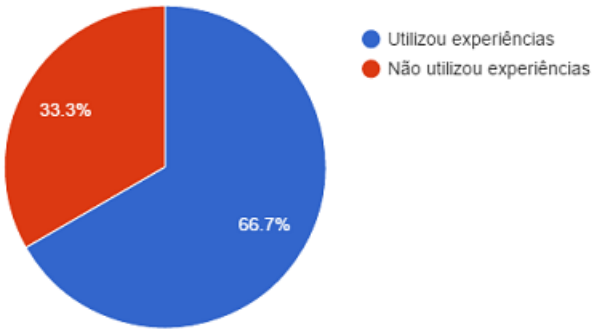

Figure 6: Gráfico de publicações que aplicaram método HCA por validação através de experiências

5. Discussão

No estudo revisado em [30] foi detectado que utilizava ambos os métodos HCA e PCA, citado anteriormente.

Dentre as principais limitações que foram citadas nos artigos, dentre aqueles que utilizavam PCA, constatou-se que para uma grande quantidade de dados, caso esta não tenha um padrão pré definido anteriormente à sua aplicação ([9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [22], [30], [31], [32], [33]), esta pode ter baixa performance em questão do tempo de processamento. Foram citadas principalmente a falta de memória física para processamento ([12] [15]) e o aparecimentos de padrões cíclicos pela quantidade alta de dados ([13], [10], [11]).

Já para as limitações dos artigos que utilizavam o método HCA foram citadas a necessidade de aplicar outros métodos para chegar a uma conclusão devido à grande quantidade de dados [34], tendo em comum características de pré definição de um padrão para ser aplicado ao conjunto de dados ([34], [35], [36], [37], [38], [39], [40], [41], [42], [30], [43], [44]).

A análise mostra claramente que há uma aplicação maior do método de PCA em distintas áreas de conhecimento, a qual está sendo a cada ano mais aplicada. Esta ainda se mostrou eficaz em conjuntos de dados que não haviam sido normalizados e não tinham ainda um padrão, como foi detectado em 100%

dos trabalhos.

6. Conclusão

Esta revisão sistemática foi realizada com o objetivo de selecionar e classificar os estudos nas diversas áreas de pesquisa que utilizam métodos HCA e PCA, da análise exploratória multivariada, tendo como finalidade a aplicação de cada método e suas limitações quanto ao conjunto de dados complexos. Desse modo, os estudos foram mapeados na Tabela 2, tais dados foram analisados e dispostos na Figura 3 e 4, as limitações discutidas na seção 4, assim como a seleção e classificação dos estudos incluídos neste estudo de revisão.

A análise do estado dos métodos estudados sugere que há uma demanda maior na aplicação do PCA diante do HCA para a descoberta de informações e validações de dados. Tal resultado pode ter relação com a ausência de padrão inicial para descoberta de conhecimentos, característica comum existente em todos os trabalhos que utilizavam o PCA. Foi detectado também que esta aplicação tem crescido a partir do ano de 2014 nas áreas de Ciências biológicas, Ciências Exatas e da Terra e Ciências da Saúde.

Esta análise também verificou que grande parte dos trabalhos validaram suas soluções com experimentos, cerca de 84,5% com método PCA e 66,7% com método HCA, revelando a necessidade de estudos que apliquem tais técnicas com dados reais.

Enfim, o presente estudo identificou dois pontos para futura pesquisa de suma importância neste campo:

- validar de modo científico a viabilidade de utilização em pares dos métodos de PCA e HCA nesta mesma ordem, a fim de dissipar as limitações encontradas; e
- averiguar o principal motivo do surgimentos de padrões cíclicos.

A partir dos pontos acima mencionados é possível concluir que é necessário aprofundar nesta pesquisa para contribuir na análise exploratória multivariada,

a fim de averiguar a eficácia de sua aplicação quanto a um conjunto de dados complexos, seja com padrões pré definidos ou não.

Referências

[1] O. Boussaid, J. Darmont, F. Bentayeb, S. Loudcher, Warehousing complex data from the web, Int. J. Web Eng. Technol. 4 (4) (2008) 408 – 433. doi:10.1504/IJWET.2008.019942.
URL http://dx.doi.org/10.1504/IJWET.2008.019942

[2] A. Cuzzocrea, A. Nucita, Enhancing accuracy and expressive power of range query answers over incomplete spatial databases via a novel reasoning approach, Data & Knowledge Engineering 70 (8) (2011) 702 – 716. doi:10.1016/j.datak.2011.03.002.

[3] M. Neto, J. Machado, G. C. Moita, An introduction analysis exploratory multivariate date, Química Nova 21 (4) (1998) 467 – 469. doi:10.1590/S0100-40421998000400016.
URL http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0100-40421998000400016&lng=en&nrm=iso&tlng=pt

[4] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, R. L. Tatham, Análise multivariada de dados - 6ed, Bookman Editora, 2009.

[5] R. Kimball, M. Ross, The Data Warehouse Toolkit, 3rd Edition, John Wiley & Sons, 2013.

[6] Z. Tan, A. Jamdagni, X. He, P. Nanda, R. P. Liu, A system for denial-of-service attack detection based on multivariate correlation analysis, IEEE Transactions on Parallel and Distributed Systems 25 (2) (2014) 447 – 456. doi:10.1109/TPDS.2013.146.

[7] J. E. Jackson, Principal components and factor analysis: Part i –principal components, Journal of Quality Technology 12 (4).
URL https://www.researchgate.net/publication/243770535_

Principal_Components_and_Factor_Analysis_Part_I-Principal_Components

[8] B. Kitchenham, S. Charters, Guidelines for performing Systematic Literature Reviews in Software Engineering, 2007.

[9] F. W. Mauldin, F. Viola, W. F. Walker, Complex principal components for robust motion estimation, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control 57 (11) (2010) 2437–2449. doi:10.1109/TUFFC.2010.1710.

[10] A. Papaioannou, S. Zafeiriou, Principal component analysis with complex kernel: The widely linear model, IEEE Transactions on Neural Networks and Learning Systems 25 (9) (2014) 1719–1726. doi:10.1109/TNNLS.2013.2285783.

[11] X. L. Li, T. Adali, M. Anderson, Noncircular principal component analysis and its application to model selection, IEEE Transactions on Signal Processing 59 (10) (2011) 4516–4528. doi:10.1109/TSP.2011.2160631.

[12] H. Zhou, G. B. Huang, Z. Lin, H. Wang, Y. C. Soh, Stacked extreme learning machines, IEEE Transactions on Cybernetics 45 (9) (2015) 2013–2025. doi:10.1109/TCYB.2014.2363492.

[13] D. Pérez-Rial, P. López-Mahía, R. Tauler, Investigation of the source composition and temporal distribution of volatile organic compounds (vocs) in a suburban area of the northwest of spain using chemometric methods, Atmospheric Environment 44 (39) (2010) 5122 – 5132. doi:http://dx.doi.org/10.1016/j.atmosenv.2010.09.005.
URL http://www.sciencedirect.com/science/article/pii/S1352231010007594

[14] M. Silvestri, A. Elia, D. Bertelli, E. Salvatore, C. Durante, M. L. Vigni, A. Marchetti, M. Cocchi, A mid level data fusion strategy

for the varietal classification of lambrusco {PDO} wines, Chemometrics and Intelligent Laboratory Systems 137 (2014) 181 – 189. doi:http://dx.doi.org/10.1016/j.chemolab.2014.06.012.
URL http://www.sciencedirect.com/science/article/pii/S0169743914001336

[15] Q. Tang, C. Bu, Y. Liu, L. Qi, Z. Yu, A new signal processing algorithm of pulsed infrared thermography, Infrared Physics & Technology 68 (2015) 173 – 178. doi:http://dx.doi.org/10.1016/j.infrared.2014.12.002.
URL http://www.sciencedirect.com/science/article/pii/S1350449514002709

[16] J. W. McIlroy, R. W. Smith, V. L. McGuffin, Assessing the effect of data pretreatment procedures for principal components analysis of chromatographic data, Forensic Science International 257 (2015) 1 – 12. doi:http://dx.doi.org/10.1016/j.forsciint.2015.07.038.
URL http://www.sciencedirect.com/science/article/pii/S0379073815003163

[17] R. Tolosana-Delgado, J. McKinley, Exploring the joint compositional variability of major components and trace elements in the tellus soil geochemistry survey (northern ireland), Applied Geochemistry (2016) –doi:http://dx.doi.org/10.1016/j.apgeochem.2016.05.004.
URL http://www.sciencedirect.com/science/article/pii/S0883292716300750

[18] J. E. Carlson, J. R. Gasson, T. Barth, I. Eide, Extracting homologous series from mass spectrometry data by projection on predefined vectors, Chemometrics and Intelligent Laboratory Systems 114 (2012) 36 – 43. doi:http://dx.doi.org/10.1016/j.chemolab.2012.02.007.
URL http://www.sciencedirect.com/science/article/pii/S016974391200041X

[19] K. Hjelle, T. Solem, L. Halvorsen, L. Åstveit, Human impact and

landscape utilization from the mesolithic to medieval time traced by high spatial resolution pollen analysis and numerical methods, Journal of Archaeological Science 39 (5) (2012) 1368 – 1379. doi:http://dx.doi.org/10.1016/j.jas.2011.12.026.
URL http://www.sciencedirect.com/science/article/pii/S0305440311004730

[20] P. Penczek, M. Kimmel, C. Spahn, Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-em images, Structure 19 (11) (2011) 1582 – 1590. doi:http://dx.doi.org/10.1016/j.str.2011.10.003.
URL http://www.sciencedirect.com/science/article/pii/S0969212611003571

[21] D. Rutledge, D. J.-R. Bouveresse, Independent components analysis with the {JADE} algorithm, TrAC Trends in Analytical Chemistry 50 (2013) 22 – 32. doi:http://dx.doi.org/10.1016/j.trac.2013.03.013.
URL http://www.sciencedirect.com/science/article/pii/S0165993613001222

[22] R. Danielsson, E. Allard, P. J. R. Sjöberg, J. Bergquist, Exploring liquid chromatography–mass spectrometry fingerprints of urine samples from patients with prostate or urinary bladder cancer, Chemometrics and Intelligent Laboratory Systems 108 (1) (2011) 33 – 48, analytical Platforms for Providing and Handling Massive Chemical Data. doi:http://dx.doi.org/10.1016/j.chemolab.2011.03.008.
URL http://www.sciencedirect.com/science/article/pii/S016974391100058X

[23] P. K. Zarzycki, J. K. Portka, Recent advances in hopanoids analysis: Quantification protocols overview, main research targets and selected problems of complex data exploration, The Journal of Steroid Biochemistry and Molecular Biology 153 (2015) 3 – 26, perspectives in Steroid Research.

doi:http://dx.doi.org/10.1016/j.jsbmb.2015.04.017.
URL http://www.sciencedirect.com/science/article/pii/S0960076015001223

[24] A. Hertrampf, R. Sousa, J. Menezes, T. Herdling, Semi-quantitative prediction of a multiple {API} solid dosage form with a combination of vibrational spectroscopy methods, Journal of Pharmaceutical and Biomedical Analysis 124 (2016) 246 – 253. doi:http://dx.doi.org/10.1016/j.jpba.2016.03.003.
URL http://www.sciencedirect.com/science/article/pii/S0731708516301212

[25] P. Liu, Z. Li, B. Li, G. Shi, M. Li, D. Yu, J. Liu, The analysis of time-resolved optical waveguide absorption spectroscopy based on positive matrix factorization, Journal of Colloid and Interface Science 403 (2013) 134 – 141. doi:http://dx.doi.org/10.1016/j.jcis.2013.03.035.
URL http://www.sciencedirect.com/science/article/pii/S0021979713002804

[26] Z. Ma, X. Song, R. Wan, L. Gao, A modified water quality index for intensive shrimp ponds of litopenaeus vannamei, Ecological Indicators 24 (2013) 287 – 293. doi:http://dx.doi.org/10.1016/j.ecolind.2012.06.024.
URL http://www.sciencedirect.com/science/article/pii/S1470160X12002658

[27] D. Shen, H. Shen, J. Marron, Consistency of sparse {PCA} in high dimension, low sample size contexts, Journal of Multivariate Analysis 115 (2013) 317 – 333. doi:http://dx.doi.org/10.1016/j.jmva.2012.10.007.
URL http://www.sciencedirect.com/science/article/pii/S0047259X12002308

[28] R. V. Haware, P. R. Wright, K. R. Morris, M. L. Hamad, Data fusion of fourier transform infrared spectra and powder x-ray

diffraction patterns for pharmaceutical mixtures, Journal of Pharmaceutical and Biomedical Analysis 56 (5) (2011) 944 – 949. doi:http://dx.doi.org/10.1016/j.jpba.2011.08.018.
URL http://www.sciencedirect.com/science/article/pii/S0731708511004560

[29] V. Singh, H. Agrawal, G. Joshi, M. Sudershan, A. Sinha, Elemental profile of agricultural soil by the {EDXRF} technique and use of the principal component analysis (pca) method to interpret the complex data, Applied Radiation and Isotopes 69 (7) (2011) 969 – 974. doi:http://dx.doi.org/10.1016/j.apradiso.2011.01.025.
URL http://www.sciencedirect.com/science/article/pii/S096980431100039X

[30] A. Menció, A. Folch, J. Mas-Pla, Identifying key parameters to differentiate groundwater flow systems using multifactorial analysis, Journal of Hydrology 472–473 (2012) 301 – 313. doi:http://dx.doi.org/10.1016/j.jhydrol.2012.09.030.
URL http://www.sciencedirect.com/science/article/pii/S0022169412008335

[31] C. Hu, J. Sepulcre, K. A. Johnson, G. E. Fakhri, Y. M. Lu, Q. Li, Matched signal detection on graphs: Theory and application to brain imaging data classification, NeuroImage 125 (2016) 587 – 600. doi:http://dx.doi.org/10.1016/j.neuroimage.2015.10.026.
URL http://www.sciencedirect.com/science/article/pii/S1053811915009362

[32] Y. Chen, T. ting Huang, H. lin Liu, D. Zhan, Multi-pose face ensemble classification aided by gabor features and deep belief nets, Optik - International Journal for Light and Electron Optics 127 (2) (2016) 946 – 954. doi:http://dx.doi.org/10.1016/j.ijleo.2015.10.179.

URL http://www.sciencedirect.com/science/article/pii/S0030402615015478

[33] S. K. Ryman, N. D. Bruce, M. S. Freund, Temporal responses of chemically diverse sensor arrays for machine olfaction using artificial intelligence, Sensors and Actuators B: Chemical 231 (2016) 666 – 674. doi:http://dx.doi.org/10.1016/j.snb.2016.03.059.
URL http://www.sciencedirect.com/science/article/pii/S0925400516303537

[34] H. D. Meng, Y. C. Song, F. Y. Song, H. T. Shen, Application research of cluster analysis and association analysis, in: Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on, 2010, pp. 597–602.

[35] N. Cao, D. Gotz, J. Sun, H. Qu, Dicon: Interactive visual analysis of multidimensional clusters, IEEE Transactions on Visualization and Computer Graphics 17 (12) (2011) 2581–2590. doi:10.1109/TVCG.2011.188.

[36] D. Karthik, K. VijayaRekha, K. Manjula, Multivariate analysis for detecting adulteration in edible oil: A review, in: Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on, 2012, pp. 272–277.

[37] M. Kurematsu, H. Fujita, A framework for integrating a decision tree learning algorithm and cluster analysis, in: Intelligent Software Methodologies, Tools and Techniques (SoMeT), 2013 IEEE 12th International Conference on, 2013, pp. 225–228. doi:10.1109/SoMeT.2013.6645670.

[38] P. V. Campregher, S. K. Srivastava, H. J. Deeg, H. S. Robins, E. H. Warren, Abnormalities of the α β t-cell receptor repertoire in advanced myelodysplastic syndrome, Experimental Hematology 38 (3) (2010) 202 – 212. doi:http://dx.doi.org/10.1016/j.exphem.2009.12.004.
URL http://www.sciencedirect.com/science/article/pii/S0301472X09004615

[39] T. Hatori, M. Sato-Ilic, A fuzzy clustering method using the relative structure of the belongingness of objects to clusters, Procedia Computer Science 35 (2014) 994 – 1002, knowledge-Based and Intelligent Information & Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, September 2014 Proceedings. doi:http://dx.doi.org/10.1016/j.procs.2014.08.185.
URL http://www.sciencedirect.com/science/article/pii/S1877050914011508

[40] K. Banas, A. M. Banas, M. Gajda, W. M. Kwiatek, B. Pawlicki, M. B. Breese, Analysis of synchrotron radiation induced x-ray emission spectra with r environment, Radiation Physics and Chemistry 93 (2013) 82 – 86, proceedings of the 11th International School and Symposium on Synchrotron Radiation in Natural Science (ISSRNS). doi:http://dx.doi.org/10.1016/j.radphyschem.2013.04.026.
URL http://www.sciencedirect.com/science/article/pii/S0969806X13002545

[41] H. Yu, Z. Liu, G. Wang, An automatic method to determine the number of clusters using decision-theoretic rough set, International Journal of Approximate Reasoning 55 (1, Part 2) (2014) 101 – 115, special issue on Decision-Theoretic Rough Sets. doi:http://dx.doi.org/10.1016/j.ijar.2013.03.018.
URL http://www.sciencedirect.com/science/article/pii/S0888613X13000868

[42] B. R. Parresol, J. H. Scott, A. Andreu, S. Prichard, L. Kurth, Developing custom fire behavior fuel models from ecologically complex fuel structures for upper atlantic coastal plain forests, Forest Ecology and Management 273 (2012) 50 – 57, assessing wildland fuels and hazard mitigation treatments in the southeastern United States. doi:http://dx.doi.org/10.1016/j.foreco.2012.01.024.

URL http://www.sciencedirect.com/science/article/pii/S0378112712000345

[43] T. Chen, N. L. Zhang, T. Liu, K. M. Poon, Y. Wang, Model-based multidimensional clustering of categorical data, Artificial Intelligence 176 (1) (2012) 2246 – 2269. doi:http://dx.doi.org/10.1016/j.artint.2011.09.003.
URL http://www.sciencedirect.com/science/article/pii/S000437021100110X

[44] D.-S. Cao, J.-H. Huang, Y.-Z. Liang, Q.-S. Xu, L.-X. Zhang, Tree-based ensemble methods and their applications in analytical chemistry, TrAC Trends in Analytical Chemistry 40 (2012) 158 – 167. doi:http://dx.doi.org/10.1016/j.trac.2012.07.012.
URL http://www.sciencedirect.com/science/article/pii/S0165993612002336