





Métodos Quantitativos Interpretativos na Educação em Ciências: Abordagens para Análise Multivariada de Dados

Interpretative Quantitative Methods in Science Education: Approaches to Multivariate Data Analysis

Matheus Monteiro Nascimento  Brasil
Estevão Antunes Júnior  Brasil
Cláudio José de Holanda Cavalcanti  Brasil
Fernanda Ostermann  Brasil

A área de pesquisa em Educação em Ciências ainda sofre influências do positivismo, materializadas no delineamento de estudos tanto quantitativos como qualitativos. Os métodos quantitativos tradicionais se mostram como fortes agentes da perspectiva positivista de pesquisa, principalmente pela ingênua interpretação de que os dados numéricos representam a realidade dos fatos. Na intenção de se afastar dessa visão, observou-se um avanço na direção do desenvolvimento dos chamados métodos mistos, principalmente quanto à análise quantitativa interpretativa, que incentiva uma variedade maior de métodos analíticos para poder melhor compreender o objeto de estudo. Assim, objetiva-se discutir, nesse artigo, o Escalonamento Multidimensional (*Multidimensional Scaling* - EMD) e a Análise de Correspondência (*Correspondence Analysis* - AC), que se configuram como possibilidades de métodos quantitativos interpretativos recorrentemente utilizados na análise multivariada de dados, além de apresentar exemplos de aplicação em pesquisas na área de Educação em Ciências. O resultado da nossa apresentação é que métodos quantitativos interpretativos devidamente fundamentados e refletidos podem se afastar do paradigma positivista que ainda permeia a área de Educação em Ciências.

Palavras-chave: Análise de Correspondência; Análise Multivariada; Escalonamento Multidimensional.

The field of research in Science Education still lives in the shadow of positivism, which hides in methodologies both quantitative and qualitative. Traditional quantitative methods are shown to be strong agents of the positivist perspective of research, mainly due to the naive interpretation that the numerical data represent the reality of the facts. In order to move away from this positivist view, there was an advance in the direction of the development of so-called mixed methods, especially in the quantitative interpretive analysis, which encourages a greater variety of analytical methods in order to better understand the object of study. Thus, the objective of this paper is to discuss

Multidimensional Scaling (EMD) and Correspondence Analysis (AC), which are configured as possibilities of interpretative quantitative methods that are routinely used in the multivariate analysis of data, as well as examples applied in research in the area of Science Education. The result of our presentation is that duly substantiated and reflected interpretative quantitative methods can move away from the positivist paradigm that still permeates the area of Science Education.

Keywords: Correspondence Analysis; Multivariate Analysis; Multidimensional Scaling.

1. Introdução

Muitas áreas do conhecimento buscam no cientificismo uma forma de se legitimar. Essa visão se tornou hegemônica no Brasil nas décadas de 50 e 60 do século passado, em que as políticas educacionais se direcionavam fortemente à percepção do filósofo norte americano John Dewey. O objetivo era proporcionar à educação um método, semelhante ao científico, que forneceria uma sequência de passos prescritos, ou seja, uma ciência da educação (Cunha, 2001).

A história mostra que as ciências naturais ganharam uma relevância muito grande, em vários momentos, por estarem representando uma suposta verdade pura e absoluta. Assim, surge o chamado *método científico* para explicar fenômenos na ciência. Tal movimento é atualmente visto como uma visão de produção de conhecimento ultrapassada, inclusive nas áreas mais tradicionais como a própria Física. Essa visão, ainda que não esteja assumidamente presente nesses discursos, principalmente os científicos, é encontrada implicitamente em referenciais teóricos e metodológicos.

Metodologicamente, a ideia de buscar uma verdade absoluta foi muito associada, ao longo do tempo, a processos de análise quantitativa de dados (Bryman, 1984). Filosoficamente, essa identificação assume que pesquisas quantitativas, nessa perspectiva, procuram encontrar nos dados, a partir do auxílio de representações estatísticas, uma suposta realidade objetiva. Isso seria feito a partir da simulação da influência de variáveis independentes sobre variáveis dependentes, buscando-se encontrar relações de causalidade. Outrossim, quando a pesquisa quantitativa é analisada sob esse viés, como se os dados representassem de fato o que o pesquisador necessita analisar, subverte-se o sentido principal da análise e se desconsidera fatores extra numéricos, que podem afetar os resultados, assim como a própria percepção e a manipulação do pesquisador. Por exemplo, em exames de larga escala como o SAEB, Prova Brasil, Pisa e ENEM há uma correlação muito forte entre o desempenho dos estudantes e a renda familiar média. Uma interpretação aligeirada poderia concluir que estudantes ricos são mais inteligentes do que alunos de baixa renda. Essa visão simplista desconsidera todos os elementos do contexto social das crianças que podem influenciar o seu desempenho escolar.

Ainda, é possível encontrar resquícios dessa crença na realidade objetiva não apenas na tradição de pesquisas quantitativas, pois as metodologias de pesquisa qualitativas podem incorporar essa filosofia a fim de atribuir maior credibilidade aos

dados não numéricos. Sobre metodologias de pesquisa qualitativa que carregam de forma subjacente aspectos desse ideário, cita-se a Teoria Fundamentada, que admite que os dados qualitativos brutos podem ser fracionados por meio de categorias e subcategorias, de forma a extrair teorias (teorias *emergem* dos dados) para se chegar a conclusões fidedignas (Age, 2011; Cho; Lee, 2014).

Não é de hoje que existe a preocupação dos pesquisadores da área da Educação em Ciências com questões metodológicas da pesquisa. Já no Encontro Nacional de Pesquisa em Educação em Ciências (ENPEC) de 2001, evento organizado pela Associação Brasileira de Pesquisa em Educação em Ciências (ABRAPEC), houve uma mesa redonda justamente para debater este tema. Alguns pontos debatidos no evento contribuem para a justificação do presente trabalho e merecem algum destaque. Medeiros (2002) ressalta a hegemonia do paradigma qualitativo nas pesquisas da área, revertendo a tendência quantitativa observada no campo educacional desde meados da década de 60. Esta reversão, contudo, levou a situação para outro extremo, ao ponto de se observar uma tendência das revistas especializadas de não aceitarem pesquisas quantitativas (Medeiros, 2002). O autor ainda manifesta o seu desacordo com a separação maniqueísta da pesquisa em ensino em dois polos estanques: o qualitativo e o quantitativo. Nesse sentido, Greca (2002) recomenda a integração das abordagens qualitativas e quantitativas, a fim de minimizar as limitações intrínsecas de cada um dos métodos. Greca (2002) destacou também que as pesquisas da área apresentam alguns problemas sobre questões metodológicas, principalmente por não oferecerem uma discussão explícita sobre a metodologia adotada ou por não realizarem uma aproximação entre o referencial teórico utilizado e a metodologia de pesquisa. Sobre as investigações com suporte em análises estatísticas, a autora salienta que parecem se limitar apenas ao levantamento de frequência de respostas a questionários ou ao uso exclusivo de médias estatísticas sem um maior detalhamento sobre os dados.

Nesta mesma linha, Carvalho, Oliveira e Rezende (2009) realizaram uma análise de 83 artigos publicados na Revista Brasileira de Pesquisa em Educação em Ciências, periódico administrado pela ABRAPEC, entre 2004 e 2008. O objetivo da pesquisa foi analisar o conteúdo específico, a temática e o tipo de abordagem metodológica presente nesses artigos. As autoras destacam que apenas um trabalho foi identificado como quantitativo, seis como qualitativo-quantitativo e 76 como qualitativos. Mais recentemente, Schneider, Fujii e Corazza (2017) fizeram um levantamento semelhante em revistas da área da educação em ciências entre os anos de 2015 e 2016. As autoras concluem, a partir da análise de 240 artigos, que nenhum adotou uma abordagem quantitativa e apenas dez por cento utilizaram uma perspectiva mista entre quali e quantitativo. Portanto, apesar da forte recomendação dos pesquisadores da área sobre a integração entre as abordagens qualitativas e quantitativas, o que se observa é a preponderância das análises qualitativas.

Na intenção de integrar estas duas abordagens, observa-se na literatura um avanço na direção do desenvolvimento dos chamados métodos mistos, que combinam análises

quantitativas e qualitativas (Brannen, 2017; Johnson; Onwuegbuzie, 2004). Portanto, no presente trabalho tem-se o interesse particular de apresentar alguns métodos quantitativos que se alinham com a chamada análise quantitativa interpretativa (Babones, 2016), a qual se encontra dentro do paradigma metodológico dos métodos mistos (Johnson, & Onwuegbuzie, 2004). Nessa visão, o pesquisador não limita sua investigação à relação causal de variáveis objetivas. Pelo contrário, a análise quantitativa interpretativa tenta se afastar da visão tradicional das análises quantitativas, na qual hipóteses são testadas para a determinação da significância estatística de resultados sobre a relação entre variáveis. Nessa perspectiva interpretativa, que historicamente foi associada com pesquisas qualitativas, utiliza-se e recomenda-se uma variedade maior de métodos analíticos com o intuito de melhor compreender a relação entre conceitos latentes (Babones, 2016). Os métodos quantitativos interpretativos, assim, se mostram como potenciais alternativas para pesquisadores que buscam ampliar o entendimento das variáveis envolvidas na investigação, podendo “enxergar” para além dos dados brutos.

Para ilustrar esta potencialidade, Babones (2016) cita o caso mais famoso de utilização de uma análise interpretativa quantitativa, antes mesmo da existência da estatística moderna. O autor se refere ao médico britânico John Snow, que em 1854 descobriu um poço contaminado por cólera em Londres, que estava causando um número expressivo de mortes, apenas a partir da distância dos moradores até o poço e a presença da doença nestes moradores (Cerdeira, & Valdivia, 2007). Apesar de não conseguir observar a cólera nem no poço nem nas vítimas, o médico foi capaz de alertar as autoridades sobre os riscos de ingestão daquelas águas contaminadas (Babones, 2016). Ou seja, a partir de uma interpretação de dados observáveis - distância dos moradores até o poço e presença ou não da doença -, foi possível inferir a existência de uma variável não observável e ainda formular uma política de intervenção.

Desde a época de John Snow até hoje, foram observados avanços tecnológicos importantes na análise computacional de dados psicológicos, sociológicos e educacionais. Isso se evidencia pela facilidade com que os computadores conseguem analisar quantidades cada vez maiores de dados. Com esse avanço, pesquisadores têm se destinado ao aprimoramento de modelos teóricos, especialmente aqueles voltados às análises multivariadas, ferramentas de destaque dentro do paradigma dos métodos mistos (Johnson; Onwuegbuzie, 2004). De um modo geral, uma análise multivariada se refere a todo método estatístico que analisa simultaneamente múltiplas medidas de algum objeto que esteja sob investigação (Hair, Black, Babin, Anderson, & Tatham, 2009).

Sabendo das dificuldades e complicações inerentes ao processo de entendimento destes métodos quantitativos mais sofisticados, e que são fundamentais para a análise interpretativa, o objetivo deste trabalho é contribuir com a minimização dessas dificuldades a partir da discussão de dois tratamentos quantitativos recorrentemente utilizados na análise multivariada de dados e sua aplicabilidade nas pesquisas da área da Educação em Ciências. São abordados o Escalonamento Multidimensional (*Multidimensional Scaling* - EMD) e a Análise de Correspondência (*Correspondence*

Analysis - AC). A importância deste trabalho reside no fato de, além de apresentar estas duas ferramentas estatísticas sofisticadas, discutir e exemplificar as potencialidades destes métodos quando inseridos no contexto da análise quantitativa interpretativa. Quer dizer, se intenta transmitir para os pesquisadores a evolução da estatística e da filosofia por traz das análises quantitativas pautadas por uma visão interpretativa, que lança luz sobre novas questões e amplia o espectro de possibilidades analíticas em relação aos problemas da área da Educação em Ciências.

Nas próximas seções, cada um dos métodos é apresentado a partir da seguinte estratégia: primeiramente um problema genérico, sem relação com pesquisas reais, é abordado a fim de ilustrar a utilização destas ferramentas analíticas. Após o detalhamento destes exemplos são destacadas algumas pesquisas recentes da área da Educação em Ciências que se apoiaram nesses métodos e que se encaixam na abordagem quantitativa interpretativa. Não se julga possível esgotar todas as possibilidades de utilização desses tratamentos estatísticos com um trabalho, pelo contrário, o propósito desse texto é oferecer uma visão introdutória do tema e apresentar exemplos do seu emprego na área da Educação em Ciências. Ademais, este texto serve como leitura inicial para aqueles pesquisadores que objetivam realizar investigações apoiadas na abordagem dos métodos quantitativos interpretativos.

É importante destacar que todas as análises apresentadas neste trabalho foram realizadas no ambiente de programação R (R Core Team, 2015).

2. Escalonamento Multidimensional

O Escalonamento Multidimensional é uma das mais simples ferramentas de análise multidimensional, inicialmente desenvolvida para mapear distâncias entre objetos (Johnson, & Wichern, 2007). Esse procedimento permite ao pesquisador visualizar em um plano bidimensional similaridades entre casos levando-se em conta certas variáveis. A ferramenta permite transformar similaridades entre percepções sobre algum tema específico em distâncias apresentadas em um espaço multidimensional. Nesse espaço, cada objeto ou evento é representado por um ponto, nos quais a distância entre os pares de pontos indica a relação de semelhança entre eles. As suas aplicações mais comuns são na área da administração (Borg; Groenen; Mair, 2017; Young, 2013) – percepção de consumidores sobre algum produto ou marca – e na área da ecologia (Cox; Cox, 2000; Dixon, 2003) – mapeamento de espécies em diferentes regiões.

Na área da Educação em Ciências, especialmente no Brasil, a utilização do EMD teve origem em pesquisas sobre mapeamento cognitivo e similaridade entre conceitos físicos (Greca; Moreira, 2001; Santos, 1978; Santos; Moreira, 1991), contudo, observa-se que mais recentemente esse procedimento vem sendo utilizado em estudos sobre as representações sociais das ciências (Hilger, 2009; Rodrigues; Borges; Pietrocola, 2018). Apesar das investigações destacadas, o potencial da aplicabilidade do EMD ainda é pouco discutido no cenário acadêmico brasileiro. Por isso, nesta seção será realizado um EMD a partir de dados genéricos, com o objetivo de deixar claro para o leitor o sentido

de utilização dessa ferramenta e posteriormente serão apresentadas investigações que se apoiaram no EMD para estudos distintos dos descritos anteriormente, a fim de ilustrar as suas múltiplas aplicações.

O Escalonamento Multidimensional está fortemente relacionado com o conceito de dissimilaridade que, por sua vez, guarda uma relação complementar com o conceito de similaridade, como mostra a relação matemática $d = 1 - s$, onde d é a dissimilaridade e s a similaridade. Assim, se $s = 1$ (dois sujeitos ou objetos idênticos segundo uma série de parâmetros), teremos $d = 0$ e vice-versa. Conforme tratado no início dessa seção, o EMD é capaz de representar em um mapa multidimensional a similaridade entre casos. Para isso, é necessário primeiramente construir a chamada matriz de dissimilaridade.

Um exemplo clássico e intuitivo de matriz de dissimilaridade para realizar um EMD é uma matriz de distâncias euclidianas entre cidades de um país (nas linhas e colunas estão os nomes das cidades, que mantivemos no idioma original). Os elementos dessa matriz seriam as distâncias entre as cidades. Assim, a “dissimilaridade” entre pares de cidades seria quantificada unicamente pela distância euclidiana entre elas. É importante salientar que o conceito de dissimilaridade não se limita ao de distância euclidiana. Existem diversas medidas de dissimilaridade na estatística e apenas uma delas é a distância euclidiana, que provavelmente é uma das fontes de confusão entre os conceitos de distância e dissimilaridade. A matriz de dissimilaridade lembra uma matriz de correlações ao inverso (correlação está mais próximo de similaridade).

Como um exemplo, a Figura 1 mostra distâncias (euclidianas) entre algumas cidades da Europa, constituindo uma típica matriz de dissimilaridade. Esses dados foram obtidos a partir do *dataset* eurodist, presente no ambiente de programação R e no qual constam as distâncias, em quilômetros, entre 21 cidades da Europa.

	Athens	Barcelona	Brussels	Calais	Cherbourg	Cologne	Copenhagen	Genève	Gibraltar	Hamburg	Hook of Holland	Lisbon	Lyon	Madrid	Marseille	Milan	Munich	Paris	Rome	Stockholm	Vienna
Athens	0	3313	2963	3175	3339	2762	3276	2610	4485	2977	3030	4532	2753	3949	2865	2282	2179	3000	817	3927	1991
Barcelona	3313	0	1318	1326	1294	1498	2218	803	1172	2018	1490	1305	645	636	521	1014	1365	1033	1460	2868	1802
Brussels	2963	1318	0	204	583	206	966	677	2256	597	172	2084	690	1558	1011	925	747	285	1511	1616	1175
Calais	3175	1326	204	0	460	409	1136	747	2224	714	330	2052	739	1550	1059	1077	977	280	1662	1786	1381
Cherbourg	3339	1294	583	460	0	785	1545	853	2047	1115	731	1827	789	1347	1101	1209	1160	340	1794	2196	1588
Cologne	2762	1498	206	409	785	0	760	1662	2436	460	269	2290	714	1764	1035	911	583	465	1497	1403	937
Copenhagen	3276	2218	966	1136	1545	760	0	1418	3196	460	269	2971	1458	2498	1778	1537	1104	1176	2050	650	1455
Genève	2610	803	677	747	853	1662	1418	0	1975	1118	895	1936	158	1439	425	328	591	513	995	2068	1019
Gibraltar	4485	1172	2256	2224	2047	2436	3196	1975	0	2897	2428	676	1817	698	1693	2185	2565	1971	2631	3886	2974
Hamburg	2977	2018	597	714	1115	460	460	1118	2897	0	550	2671	1159	2198	1479	1238	805	877	1751	949	1155
Hook of Holland	3030	1490	172	330	731	269	269	895	2428	550	0	2280	863	1730	1183	1098	851	457	1683	1500	1205
Lisbon	4532	1305	2084	2052	1827	2290	2971	1936	676	2671	2280	0	1178	668	1762	2250	2507	1799	2700	3231	2937
Lyon	2753	645	690	739	789	714	1458	158	1817	1159	863	1178	0	1281	320	328	724	471	1048	2108	1157
Madrid	3949	636	1558	1550	1347	1764	2498	1439	698	2198	1730	668	1281	0	1157	1724	2010	1273	2097	3188	2409
Marseille	2865	521	1011	1059	1101	1035	1778	425	1693	1479	1183	1762	320	1157	0	618	1109	792	1011	2428	1363
Milan	2282	1014	925	1077	1209	911	1537	328	2185	1238	1098	2250	328	1724	618	0	331	856	586	2187	898
Munich	2179	1365	747	977	1160	583	1104	591	2565	805	851	2507	724	2010	1109	331	0	821	946	1754	428
Paris	3000	1033	285	280	340	465	1176	513	1971	877	457	1799	471	1273	792	856	821	0	1476	1827	1249
Rome	817	1460	1511	1662	1794	1497	2050	995	2631	1751	1683	2700	1048	2097	1011	586	946	1476	0	2707	1209
Stockholm	3927	2868	1616	1786	2196	1403	650	2068	3886	949	1500	3231	2108	3188	2428	2187	1754	1827	2707	0	2105
Vienna	1991	1802	1175	1381	1588	937	1455	1019	2974	1155	1205	2937	1157	2409	1363	898	428	1249	1209	2105	0

Figura 1. Matriz de dissimilaridade indicando as distâncias euclidianas entre cidades da Europa

Fonte: Próprios autores (Dados: *The Cambridge Encyclopedia*).

Nesse caso, os elementos da diagonal dessa matriz são nulos, pois representam a distância de uma cidade até ela própria (em outras palavras, isso representa uma total similaridade). Quanto maior for a distância entre duas cidades quaisquer, maior sua dissimilaridade. A maior dissimilaridade, portanto, é observada entre Lisboa (LIS) e Atenas (ATH), o que significa que são as cidades mais distantes em termos geográficos. O processo de EMD consiste, nesse momento, em reproduzir a partir dessa matriz de dissimilaridade e de algoritmos específicos uma representação pictórica dessas dissimilaridades que regeneraria as posições dessas cidades no mapa geográfico. Dessa forma, o EMD resultará em um gráfico onde cada ponto representará uma cidade.

A Figura 2 apresenta o mapa resultante do EMD realizado a partir da matriz de dissimilaridades da Figura 1, com o uso do pacote *smacof* (De Leeuw, & Mair, p., 2009) do R. Uma simples inspeção mostra que a posição das cidades no mapa resultante do EMD guarda relativa semelhança com as posições reais observadas em um mapa geográfico do continente europeu.

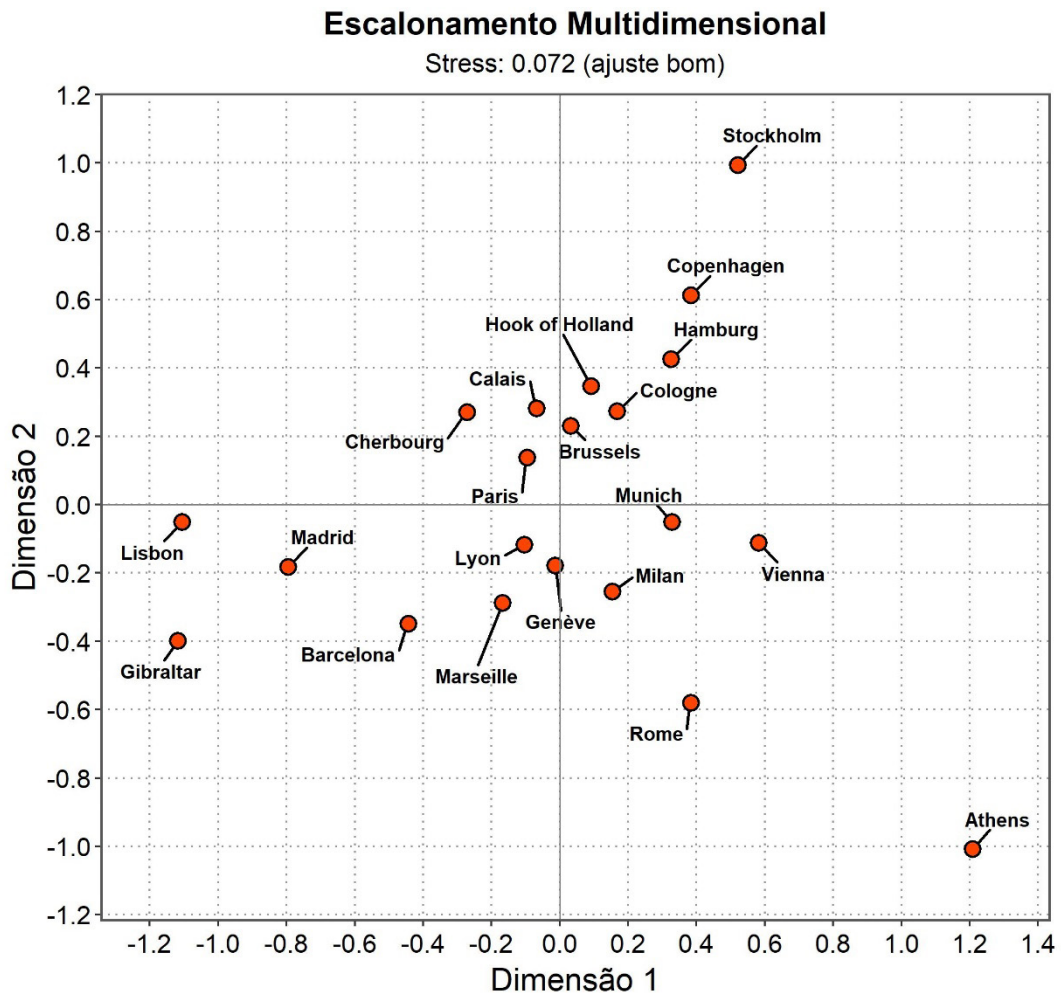


Figura 2. Mapa obtido a partir do EMD realizado na matriz de dissimilaridades na Figura 1. O mapa em geral se chama *mapa de similaridades* por permitir associar similaridades à proximidade entre os pontos. Aqui, porém, o mapa reproduz aproximadamente a distância entre as cidades e a posição relativa das mesmas (compare com um mapa geográfico)

De forma idêntica ao que foi observado na matriz de dissimilaridades, as cidades de Lisboa e Atenas são as mais distantes do mapa (valores marcados em azul na matriz da Figura 1). De maneira contrária, as duas cidades mais próximas no mapa e que logicamente possuem menor dissimilaridade, são as cidades de Lyon e Genève (valores marcados em vermelho na mesma matriz). Deve-se enfatizar que a quantidade denominada *stress*, representada sinteticamente por σ (Borg et al., 2017, p. 23; De Leeuw; Mair, 2009, seção 3.1) dá uma medida de quão bom foi o ajuste dos pontos no mapa. Pode-se usar o seguinte critério (Levshina, 2015, p. 341): $\sigma \geq 0,2$ (ajuste pobre); $0,1 \leq \sigma < 0,2$ (ajuste razoável); $0,05 \leq \sigma < 0,1$ (ajuste bom) e $\sigma \leq 0,05$ (ajuste excelente).

Esse tipo de análise, em que as dissimilaridades são propriamente as distâncias euclidianas, não é adequado para a maioria das pesquisas, incluindo as investigações da área da Educação em Ciências. A dissimilaridade não é imediatamente mensurável, devendo ser estimada seguindo critérios qualitativos e transformada em resultado numérico por algum modelo matemático. Dessa forma, para essas investigações, a construção de uma matriz de dissimilaridades deixa de ser um evento trivial como ilustrado na Figura 1. Para que isso fique mais claro, vamos construir uma matriz de dissimilaridades a partir de dados brutos, como os que são obtidos frequentemente em questionários investigativos do campo educacional, em que os respondentes podem optar em uma escala do tipo *Likert* (Likert, 1932). Como exemplo, foram utilizados dados de um questionário sobre percepção em relação à Ciência e à Tecnologia aplicados em cidadãos europeus no início dos anos 90 (Reif; Melich, 1995). Foram selecionadas sete perguntas com quatro opções de resposta cada, desde de discordo fortemente até concordo fortemente. Esses dados são encontrados no pacote *ltm* (Rizopoulos, 2006), também no R.

No intento de simplificar as análises, foram selecionados os primeiros 30 respondentes desse questionário (denotados por R1 a R30), de um total de 392. As respostas desses respondentes estão ilustradas na parte superior da Figura 3 - e abaixo estão as perguntas referentes a cada coluna. Para efeito de exercício será proposta a seguinte questão de pesquisa: qual a similaridade entre os respondentes do presente questionário? No intento de responder tal pergunta será realizado um EMD.

A escala das respostas é a seguinte: 1 (discordo fortemente), 2 (discordo), 3 (concordo) e 4 (concordo fortemente). Conforme visto, para a elaboração do mapa resultante do EMD, é necessário primeiramente construir a matriz de dissimilaridades a partir dos dados brutos. Construir uma matriz nesse formato significa calcular a dissimilaridade entre todos os pares de respondentes.

Embora seja factível calculá-la à mão, isso seria inviável para esses dados, pois levaria a uma matriz, acarretando em 450 dissimilaridades para calcular, já que a matriz é simétrica. Mesmo que individualmente o cálculo seja simples, isso não é prático. A partir da tabela com os dados brutos, que são as respostas dos diferentes respondentes - dispostos nas linhas da tabela da Figura 3(a) -, calcula-se com o R a matriz de dissimilaridades – aqui os sujeitos são os respondentes (que queremos comparar entre si) e os atributos são as respostas aos itens do questionário. Para calcular as dissimilaridades, é necessário uso de um modelo matemático que se proponha a quantificá-la. Existem muitos tipos de modelos para cálculo de dissimilaridades, cada um adequado a um tipo de dados. Para dados categóricos (nesse caso ordinais, pois é uma escala tipo *Likert*) a distância euclidiana não é a medida de dissimilaridade mais adequada. Para esse tipo de dados é possível usar a dissimilaridade de *Bray-Curtis* (Quinn, & Keough, 2002, p. 414), que pode ser facilmente obtida com o R, por exemplo, por meio do pacote *vegan* (Oksanen et al., 2019).

Existem medidas de dissimilaridade melhores para dados categóricos ou ordinais, porém bem mais complexas do que a dissimilaridade de *Bray-Curtis*, que adotaremos nesse exemplo por questão de simplicidade. Embora o R faça o cálculo de todas as dissimilaridades entre os pares e construa a matriz sem maiores problemas, como o objetivo do presente trabalho é deixar o mais acessível possível para o leitor a utilização dessas ferramentas, será calculada agora a dissimilaridade entre os dois primeiros respondentes. A dissimilaridade de *Bray-Curtis* é dada pela seguinte equação (Johnson; Myatt, 2014):

$$d_B(i, j) = \frac{\sum_{k=1}^n |x_{ik} - x_{jk}|}{\sum_{k=1}^n x_{ik} + \sum_{k=1}^n x_{jk}} \quad (1)$$

Na expressão acima, temos:

- $d_B(i, j)$ → dissimilaridade de *Bray-Curtis* entre os respondentes i e j (linhas).
- x_{ik} → valor atribuído pelo respondente i (linha) à pergunta k (coluna).
- x_{jk} → valor atribuído pelo respondente j (linha) à pergunta k (coluna).

Utilizando a equação (1) e os valores atribuídos pelos respondentes 1 e 2 – retirados da tabela superior da Figura 3 – se percebe que a dissimilaridade entre eles é de 0,09. O cálculo dessa dissimilaridade e para outros respondentes está representada na tabela inferior da Figura 3. Evidentemente a matriz de dissimilaridade é simétrica e os elementos da diagonal principal são nulos (dissimilaridade entre um sujeito e ele próprio). Nesse caso, essas quantidades estimam o quão distinta foi a sequência de respostas de um respondente em relação a outro.

Após a construção da matriz de dissimilaridades, o EMD pode ser rodado e o mapa de dissimilaridades construído. Convém lembrar que a hipotética questão de pesquisa é determinar a similaridade entre os trinta respondentes do questionário sobre percepção de Ciência e Tecnologia, ou seja, como os respondentes se agrupam ou se antagonizam em termos das respostas a cada item do questionário. A Figura 4

apresenta o mapa obtido a partir do EMD. A interpretação é muito semelhante a que foi realizada no mapa das distâncias euclidianas das cidades europeias. Porém, agora a distância no mapa indica as dissimilaridades entre os respondentes. É possível notar que quatro respondentes são totalmente similares entre si, como é o caso dos respondentes R28 e R10 e dos respondentes R27 e R30. Visualmente eles estão todos sobrepostos no mapa. Isso ocorre porque as respostas atribuídas aos sete itens do questionário foram rigorosamente as mesmas (resultando em dissimilaridade nula, marcada com vermelho na tabela inferior da Figura 3). O exame detalhado das respostas atribuídas permite, ainda, inferir o que torna esses indivíduos semelhantes, quer dizer, responderam os itens de forma similar. Esse é o caso dos respondentes R8 e R22, R18 e R28, R18 e R10 e outros.

Pode-se também inferir quais indivíduos responderam os itens do questionário de forma mais antagônica, discordando mais entre si. Os respondentes “mais antagônicos” são R24 e R29, pois apresentam a maior dissimilaridade entre si (marcada em azul na tabela inferior da Figura 3) – tal característica é bem reproduzida no mapa da Figura 4, pois nota-se que R24 e R29 são os pontos mais distantes entre si. Além disso, os respondentes R5 e R24 e R21 apresentam atitudes discordantes sobre Ciência e Tecnologia em relação aos demais respondentes considerados, pois aparecem isolados no mapa, afastados do restante do grupo.

Após a construção da matriz de dissimilaridades, o EMD pode ser rodado e o mapa de dissimilaridades construído. Convém lembrar que a hipotética questão de pesquisa é determinar a similaridade entre os trinta respondentes do questionário sobre percepção de Ciência e Tecnologia, ou seja, como os respondentes se agrupam ou se antagonizam em termos das respostas a cada item do questionário. A Figura 4 apresenta o mapa obtido a partir do EMD. A interpretação é muito semelhante a que foi realizada no mapa das distâncias euclidianas das cidades europeias. Porém, agora a distância no mapa indica as dissimilaridades entre os respondentes. É possível notar que quatro respondentes são totalmente similares entre si, como é o caso dos respondentes R28 e R10 e dos respondentes R27 e R30. Visualmente eles estão todos sobrepostos no mapa. Isso ocorre porque as respostas atribuídas aos sete itens do questionário foram rigorosamente as mesmas (resultando em dissimilaridade nula, marcada com vermelho na tabela inferior da figura 3). O exame detalhado das respostas atribuídas permite, ainda, inferir o que torna esses indivíduos semelhantes, quer dizer, quais itens responderam de forma similar. Esse é o caso dos respondentes R8 e R22, R18 e R28, R18 e R10 e outros.

Pode-se também inferir quais indivíduos responderam os itens do questionário de forma mais antagônica, discordando mais entre si. Os respondentes “mais antagônicos” são R24 e R29, pois apresentam a maior dissimilaridade entre si (marcada em azul na tabela inferior da Figura 3) – tal característica é bem reproduzida no mapa da Figura 4, pois nota-se que R24 e R29 são os pontos mais distantes entre si. Além disso, os respondentes R5 e R24 e R21 apresentam atitudes discordantes sobre Ciência e Tecnologia em relação aos demais respondentes considerados, pois aparecem isolados no mapa, afastados do restante do grupo.

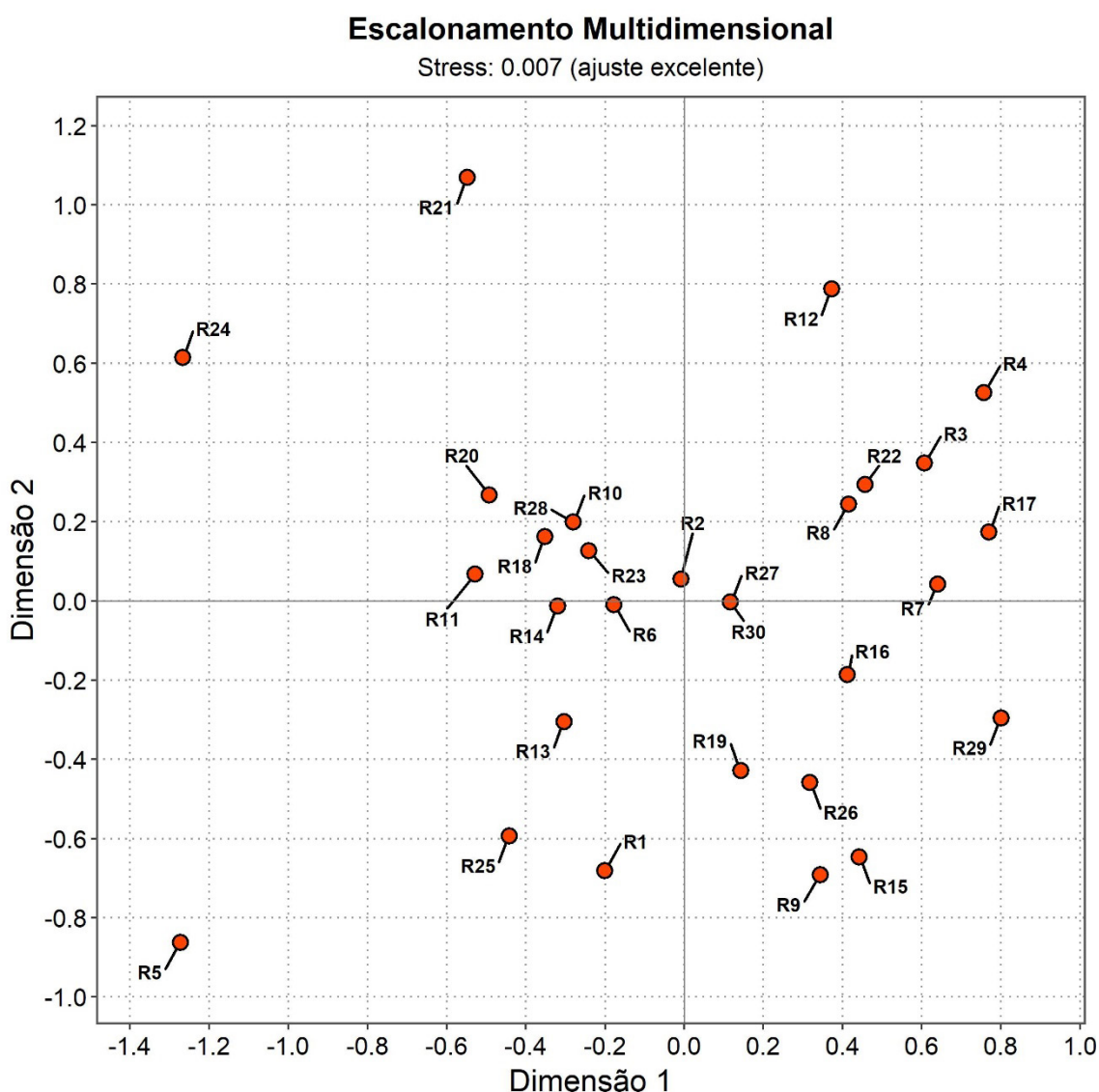


Figura 4. Mapa obtido no EMD a partir da matriz de dissimilaridades calculada com os dados do questionário sobre percepção em relação à Ciência e à Tecnologia

A despeito de ser um exemplo simples, podemos perceber que a aplicação do Escalonamento Multidimensional em pesquisas da área da Educação em Ciências se configura como uma alternativa viável e interessante quando queremos inferir relações entre sujeitos (nesse exemplo analisados pelos respectivos padrões de resposta ao questionário) difíceis de serem depreendidos diretamente dos dados brutos. Ademais, na área da Educação em Ciências são muito comuns as investigações baseadas em dados cujas respostas são ordinais, oriundas de questionários. A análise desses questionários, quando realizada a partir de métodos tradicionais como o qui-quadrado, apenas indica a associação entre as categorias, mas não diz nada sobre a relação entre os respondentes do questionário. O Escalonamento Multidimensional é capaz fazer tal associação, como ilustramos com os exemplos acima. Evidentemente, o EMD por si apenas evidencia similaridades e diferenças entre respondentes, não permitindo obter razões por essas

diferenças. Para ir mais a fundo na questão dessas diferenças e semelhanças, esse método pode ser usado de forma articulada com metodologias de pesquisa de natureza qualitativa, em um cenário investigativo fundamentado em referenciais teóricos sólidos.

2.1 O EMD em uma pesquisa recente da área

Como exemplo de pesquisa recente da área que se apoiou no EMD para determinar a similaridade entre dissertações de um curso de mestrado profissional (MP) em Ensino de Física, pode-se citar o trabalho de Nascimento (2016). O objetivo do autor foi analisar a similaridade entre 90 dissertações a partir da escolha dos referenciais teóricos. Os dados brutos para a realização da análise foram obtidos de uma tabela cujas linhas eram as dissertações (T1, T2, T3...) e as colunas os referenciais teóricos adotados nos respectivos trabalhos (Ausubel, Vygotsky, Piaget, entre outros). Se uma determinada dissertação havia se apropriado de uma abordagem freireana como referencial teórico, por exemplo, então na coluna Freire era assinalado o número 1, deixando 0 para todas as demais colunas. A similaridade investigada concentrou-se apenas nos autores adotados e tendem a aumentar quando dois trabalhos adotavam pelo menos um autor em comum, uma vez que alguns trabalhos contavam com mais de um autor como referencial teórico. Considerando que os dados são binários, nesse caso específico a dissimilaridade de *Jaccard* (Johnson, & Myatt, 2014) (mais adequada nesse caso) foi calculada e não a de *Bray-Curtis* como no exemplo anterior.

Por serem binários e indicarem presença/ausência (ou sim/não), esses dados são caracterizados como categóricos. No entanto, podem ser caracterizados também como dados ordinais, pois há uma ordenação natural: se um dado autor está presente como referencial em um certo trabalho (codificado como 1) se pode argumentar que isso o coloca em uma hierarquia superior aos que não estão (codificados com 0). Pela natureza dos dados, a melhor visualização ocorreu para três dimensões. Esse tipo de medida de dissimilaridade também pode ser calculado pelo pacote *vegan* do R.

A Figura 5 mostra o mapa resultante do EMD realizado na matriz de dissimilaridades calculada a partir dos dados sobre presença/ausência de referenciais teóricos nas dissertações de um MP. Cada ponto no mapa indica uma dissimilaridade, sendo que a coloração revela a posição na terceira dimensão. Além disso, quanto maior o tamanho do ponto, mais autores aquele trabalho utilizou como referencial teórico. A linha pontilhada vermelha delimita os grupos de trabalhos totalmente similares, sendo indicados os autores que constituem os referenciais teóricos desses grupos de trabalhos. Há claramente a formação de um grupo principal, do qual estão afastados os trabalhos T30 e os trabalhos do grupo T10-T13-T26. Há 11 trabalhos que adotam apenas Vygotsky e 48 no total que adotam o mesmo autor, ou seja, há 37 trabalhos que adotam Vygotsky entre os autores e com os quais esses 11 têm alguma similaridade. Resulta que os 11 trabalhos que adotam apenas Vygotsky guardam muito mais similaridade com os demais do que o grupo dos três que adotam Piaget – exatamente como um “efeito de atração” oriundo dessas similaridades.

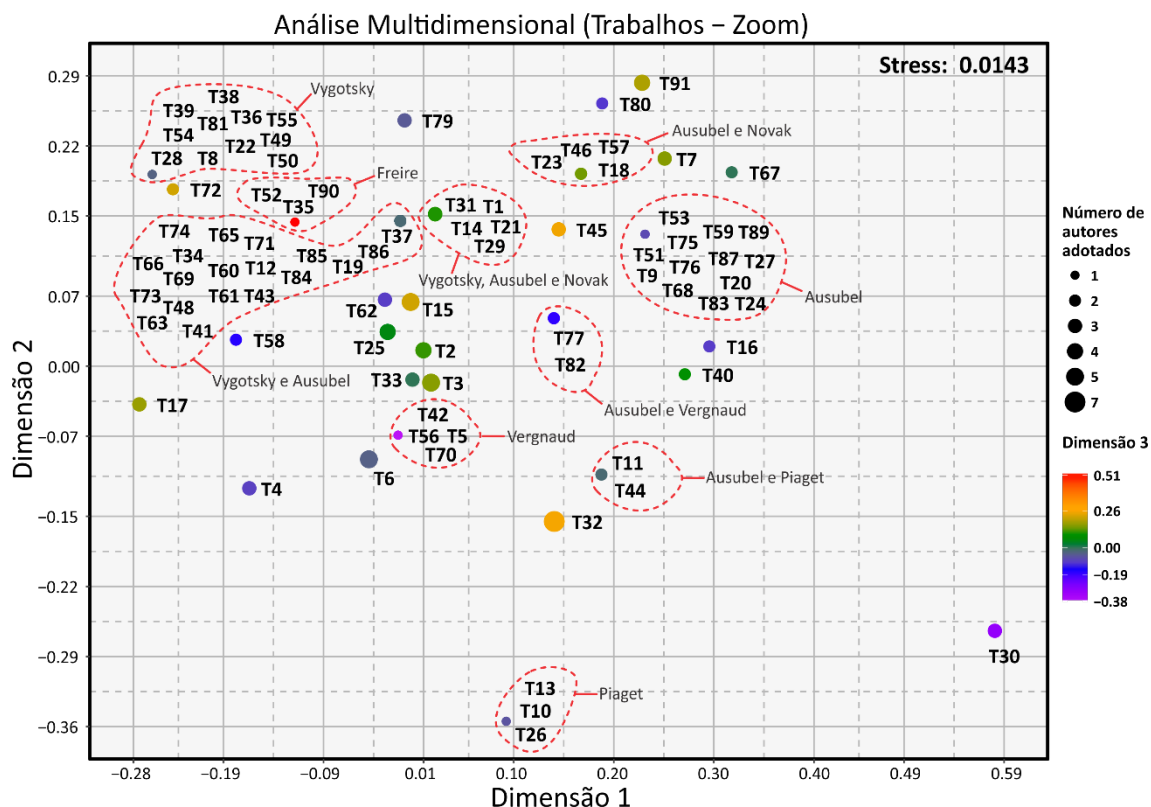


Figura 5. Análise de similaridade entre trabalhos de dissertação de um curso de Mestrado Profissional em Ensino de Física

Fonte: Nascimento (2016).

O mesmo raciocínio pode ser usado para explicar por que os trabalhos que adotam apenas Ausubel estão também mais próximos do grupo principal do que os que adotam apenas Piaget – 13 adotam apenas Ausubel e do total de 62 que adotam o mesmo autor, 49 adotam Ausubel entre os autores e estão distribuídos no grupo principal. Os trabalhos T35, T52 e T90 adotam apenas Freire, T5, T42, T56 e T70 adotam apenas Vergnaud e T64 e T78 não têm referencial teórico definido. Vemos no mapa que trabalhos que adotam apenas um autor tendem a se posicionar mais nas extremidades do mapa com exceção de T5, T42, T56 e T70, que estão próximos de T6 (que adota Vygotsky, Ausubel, Piaget, Vergnaud e Bruner) e T4 (que adota Vygotsky, Piaget e Vergnaud). Esse é um exemplo de como a escala referente à dimensão 3 é importante, pois T6 parece mais próximo do grupo T5-T42-T56-T70 do que T4. Porém, a cor de T6 mostra que ele está mais próximo do plano do que estão o referido grupo e T4. O afastamento no mapa entre o grupo T5-T42-T56-T70 e T6 é 0,346 e entre esse mesmo grupo e T4 é 0,330, ligeiramente inferior. O que o EMD forneceu foram alguns indícios sobre a forma como os referencias teóricos estavam sendo encarados no programa em questão, elementos que muito dificilmente poderiam ser identificados a partir de uma simples inspeção dos dados.

3. Análise de Correspondência

De acordo com Hair et al. (2009), a “Análise de Correspondência (AC) proporciona uma representação multivariada de interdependência para dados expressos por variáveis categóricas que outros métodos não permitem” (p. 20). Em outras palavras, o objetivo da AC é visualizar em um mapa de dimensões reduzidas a associação entre variáveis categóricas. Essa técnica permite determinar o grau de associação entre linhas e colunas de uma tabela. Em termos práticos, existe mais de um tipo de AC, mas todos podem ser divididos basicamente em dois grandes grupos, são eles: a Análise de Correspondência Simples (ACS – usada para estudar associações entre duas variáveis categóricas); e a Análise de Correspondência Múltipla (ACM – uma generalização da ACS para estudar mais associações entre múltiplas variáveis categóricas) (Greenacre, 2017). O objetivo primordial da ACM é uma *redução de dimensão*, ou seja, visualizar dados de natureza multidimensional em um número reduzido de dimensões (usualmente duas ou três, ou seja, em um espaço plano ou tridimensional, de forma que reproduza bem a variância total da tabela). Assim, análises no estilo da ACM são técnicas de *data-mining* que não visam necessariamente uma generalização a uma gama ampla de contextos, mas explicitar padrões ou associações importantes que possam ser articuladas a análises de natureza quantitativa ou qualitativa.

Essa técnica tem a sua origem matemática no trabalho de Hirschfeld (1935) e, desde então, tem sido utilizada em diferentes contextos, como na ecologia e psicologia. Por facilitar a análise da associação entre variáveis categóricas, a AC se tornou uma importante ferramenta para as investigações das áreas das ciências humanas, da saúde e sociais (Greenacre; Blasius, 1994; Ferreira, 2003), especialmente a partir da década de sessenta com os trabalhos do sociólogo francês Pierre Bourdieu (De Nooy, 2003). Na área da Educação em Ciências, por outro lado, a AC é também uma técnica que ainda não faz parte do repertório metodológico dos seus pesquisadores, guardadas algumas poucas exceções (Adamuti-Trache; Andres, 2008; Andres; Adamuti-Trache, 2008; Nascimento; Cavalcanti; Ostermann, 2017; Shanka; Quintal; Taylor, 2006).

A Análise de Correspondência (AC) tem como dados de entrada tabelas de contingência (tabelas cruzadas que mostram a distribuição de frequências das ocorrências comuns entre as diversas categorias das variáveis). Como exemplo, a Tabela 1 é a tabela de contingência de dados hipotéticos de uma pesquisa sobre a frequência com que grupos de pessoas fumam cigarro. Esses dados foram adaptados de Greenacre (1984). Essa tabela contém cinco linhas (grupos de pessoas) e quatro colunas (categorias de tabagismo), em que aparecem as frequências de categorias de tabagismo para cada grupo de funcionários em uma organização fictícia. Por exemplo, 25 pessoas do Grupo C afirmam não fumar, enquanto que 13 pessoas do Grupo D afirmam fumar muito. Essas quantidades são chamadas de valores observados. A Análise de Correspondência permite a visualização da associação entre as categorias dessas duas variáveis, que são o grupo de pessoas e as categorias de tabagismo.

Na Tabela 1, a última linha (PC) e a última coluna (PL) indicam, respectivamente,

a proporção de ocorrência de cada coluna (categorias de tabagismo) e linha (grupos de pessoas). Por exemplo, a proporção de pessoas que afirmam nunca fumar (categoria *Nada*) é dada pela razão entre a soma das frequências na coluna correspondente e o número total de casos, ou seja, $(4+4+25+18+10)/193 \approx 0,316062$. Esse número pode ser entendido como a probabilidade de ocorrência desse grupo dentro da tabela, quer dizer, a probabilidade de uma pessoa afirmar que não fuma dentro da população total investigada é de aproximadamente 31,6 por cento. Na Tabela 2 são apresentados os valores esperados para as frequências, com precisão de duas casas decimais. Essa grandeza, o valor esperado de cada caso, pode ser facilmente calculada a partir dos valores observados. Para isso, basta multiplicar a proporção de ocorrência da linha pela proporção de ocorrência da coluna e pelo número total de casos. Para o caso de não fumantes dentro do Grupo C, por exemplo, o valor esperado é dado por $0,264249 \times 0,316062 \times 193 \approx 16,12$.

Tabela 1. Tabela de Contingência

	Nada	Pouco	Médio	Muito	PC
Grupo A	4	2	3	2	0.056995
Grupo B	4	3	7	4	0.093264
Grupo C	25	10	12	4	0.264249
Grupo D	18	24	33	13	0.455959
Grupo E	10	6	7	2	0.129534
PL	0.316062	0.233161	0.321244	0.129534	

Tabela 2. Valores Esperados

	Nada	Pouco	Médio	Muito
Grupo A	3.48	2.56	3.53	1.42
Grupo B	5.69	4.20	5.78	2.33
Grupo C	16.12	11.89	16.38	6.61
Grupo D	27.81	20.59	28.27	11.4
Grupo E	7.90	5.83	8.03	3.24

Tabela 3. Resíduos padronizados para dados de fumantes de uma empresa fictícia

	Nada	Pouco	Médio	Muito
Grupo A	0.28	-0.35	-0.28	0.48
Grupo B	-0.71	-0.58	0.51	1.09
Grupo C	2.21	-0.55	-1,08	-1.01
Grupo D	-1.86	0.77	0.89	0.47
Grupo E	0.75	0.07	-0.36	-0.69

A Tabela 3 mostra os resultados mais importantes para a realização da Análise de Correspondência, que são os resíduos padronizados. O resíduo padronizado R_p é definido como

$$R_p = \frac{V_o - V_e}{\sqrt{V_e}}, \quad (2)$$

onde:

- $V_o \rightarrow$ é o valor observado;
- $V_e \rightarrow$ é o valor esperado.

O resíduo padronizado informa para cada categoria cruzada o desvio em relação ao valor esperado em unidades da raiz quadrada do valor esperado. Ou seja, ele indica o quanto que a variável observada se afastou do valor esperado. Quanto mais baixo (alto) for o módulo do resíduo, menor (maior) será a associação entre as categorias. Essa matriz com os resíduos padronizados é peça chave para a construção do mapa com a visualização da associação entre as variáveis a partir da AC. Matematicamente, a AC se baseia na decomposição em valores singulares da matriz dos resíduos para representar graficamente as linhas e as colunas da tabela de contingência como pontos em espaços vetoriais de dimensões reduzidas (Souza, Bastos, & Vieira, 2010). Essa decomposição é realizada a partir dos resíduos padronizados e será detalhada a seguir.

A decomposição em valores singulares (*Singular Value Decomposition – SVD*) é uma técnica da álgebra linear utilizada para fatorar matrizes (Anton, & Busby, 2006). A SVD decompõe uma matriz num produto dos fatores de outras três matrizes. Assim, seja \mathbf{C} uma matriz real ou complexa, então existem duas matrizes ortogonais unitárias \mathbf{U} e \mathbf{V} ; e uma matriz diagonal $\mathbf{\Sigma}$ tais que:

$$\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (3)$$

onde \mathbf{V}^T é a matriz transposta de \mathbf{V} . Os valores da diagonal de $\mathbf{\Sigma}$ são chamados de valores singulares¹, expressos como um vetor σ e correspondem à raiz quadrada dos autovalores λ da matriz \mathbf{C} .

No caso da Análise de Correspondência, a matriz é simplesmente a razão entre a matriz dos resíduos padronizados e a raiz quadrada do número total de casos, contados a partir da tabela de contingência (Tabela 1). As coordenadas dos pontos do mapa x_{linha} e x_{coluna} da AC são obtidas das matrizes \mathbf{U} e \mathbf{V} a partir das relações (Greenacre, 2017):

$$x_{\text{linha}} = \frac{\mathbf{U}\sigma}{\sqrt{PL}} \quad (4)$$

e

$$x_{\text{coluna}} = \frac{\mathbf{V}\sigma}{\sqrt{PC}} \quad (5)$$

¹ Para um maior detalhamento sobre a decomposição em valores singulares, sugere-se os seguintes links: http://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm e <http://web.cs.iastate.edu/~cs577/handouts/svd.pdf>

Portanto, o problema matemático de uma AC é a determinação das matrizes U e V a partir da matriz de entrada C . Para realizar essas operações, é preciso levar em consideração que $CC^T = V\Sigma^T\Sigma V^T$ e que $CV = U\Sigma$. É importante destacar que os cálculos para a obtenção dessas matrizes se complexificam muito a partir do aumento do número de categorias envolvidas. Dessa forma, são necessários métodos numéricos para efetuar tais operações complexas. O ambiente de programação R tem mais essa vantagem, pois nele se realizam essas operações mesmo que o número de categorias aumente substancialmente. O pacote *ca* (Nenadic, & Greenacre, 2007) do R é bastante completo e permite que se compute a AC e mais de um tipo de ACM.

Como foi apresentado, a Análise de Correspondência busca estudar se uma dada categoria se relaciona apreciavelmente com outra (ou outras). A representação gráfica é mostrada em geral em um mapa bidimensional por meio de segmentos de reta que iniciam na origem (alguns com orientação, para referência), de forma que o seu comprimento no mapa indica o quanto de informação aquela categoria agrega a este mapa. Além disso, quanto maiores o cosseno do ângulo entre dois segmentos e seus comprimentos, maior a associação entre ambos. Assim, segmentos longos orientados de forma que existe um ângulo pequeno um em relação ao outro estão fortemente associados. Segmentos longos orientados a um ângulo maior do que 90 graus (cosseno negativo) são antiassociados (anticorrelacionados, sendo essa anticorrelação máxima se o ângulo for próximo de 180 graus, ou seja, oposição máxima). No caso do ângulo ser próximo a 90 graus, as categorias possuem muito pouca associação entre si. Caso a variância total explicada pela visualização não seja muito alta (abaixo de 80 por cento) e os segmentos possuam comprimentos pequenos no mapa, a interpretação das associações envolvendo essas categorias deve ser feita com cuidado. Caso a variância total explicada pelo mapa seja alta (acima de 80 por cento), segmentos curtos indicam que a categoria não se associa (ou antiassocia) com as demais. Em outras palavras, isso indica que as ocorrências de categorias representadas por segmentos curtos não se desviam muito do valor esperado.

A Figura 6 mostra a AC resultante dos dados da tabela de contingência - Tabela 1. Ela indica a associação entre o nível de consumo de tabaco e grupos de trabalhadores de uma empresa fictícia. A qualidade da visualização é dada pela alta variância explicada pelas suas dimensões - 99,51 por cento -, ou seja, um mapa de muito alta fidelidade, reproduzindo visualmente muito bem os resíduos da Tabela 3. Foi utilizado como variável de referência o consumo de tabaco, sendo representado por setas na cor vermelha.

Os cinco grupos de trabalhadores foram posicionados no mapa e destacados por setas na cor preta. Lembrando que quanto mais próximo de zero for o ângulo entre dois segmentos, mais forte é a associação.

É importante entender que a AC mede associações *relativas*. Por exemplo, da visualização da Figura 6 é possível afirmar que os integrantes do Grupo C se caracterizam mais por não fumar *do que os demais*. Ou seja, dentre todos os grupos esse é o que mais supera o valor esperado nesse quesito, como mostra a Tabela 3 - o resíduo padronizado é 2,21. Os participantes do Grupo B se caracterizam mais do

que os outros por fumar muito (resíduo padronizado 1,91). É de se esperar, com isso, que o Grupo C esteja anticorrelacionado com a categoria de consumo *Muito*, o que se verifica pelo ângulo maior do que 90 graus entre seus segmentos. O mesmo vale para a anticorrelação entre o Grupo B e a categoria *Nada*. Vale ainda uma observação sobre o Grupo A, que praticamente não se associa fortemente a nenhuma categoria de consumo, o que é revelado pelos ângulos entre o grupo e as categorias *Nada* e *Muito*, ou pelos resíduos pequenos para esse mesmo cruzamento de atributos. O grupo D é o que mais se afasta do valor esperado na categoria de consumo *Médio*, sendo também o que mais se anticorrelaciona com a categoria *Nada*. Assim, AC permite inferir associações entre os grupos (entes) e alguns atributos (categoria de consumo), que ajuda a entender o comportamento dos grupos e relações entre eles segundo seus hábitos de consumo. A seguir será apresentada uma pesquisa feita na área de Educação em Ciências com esse método.

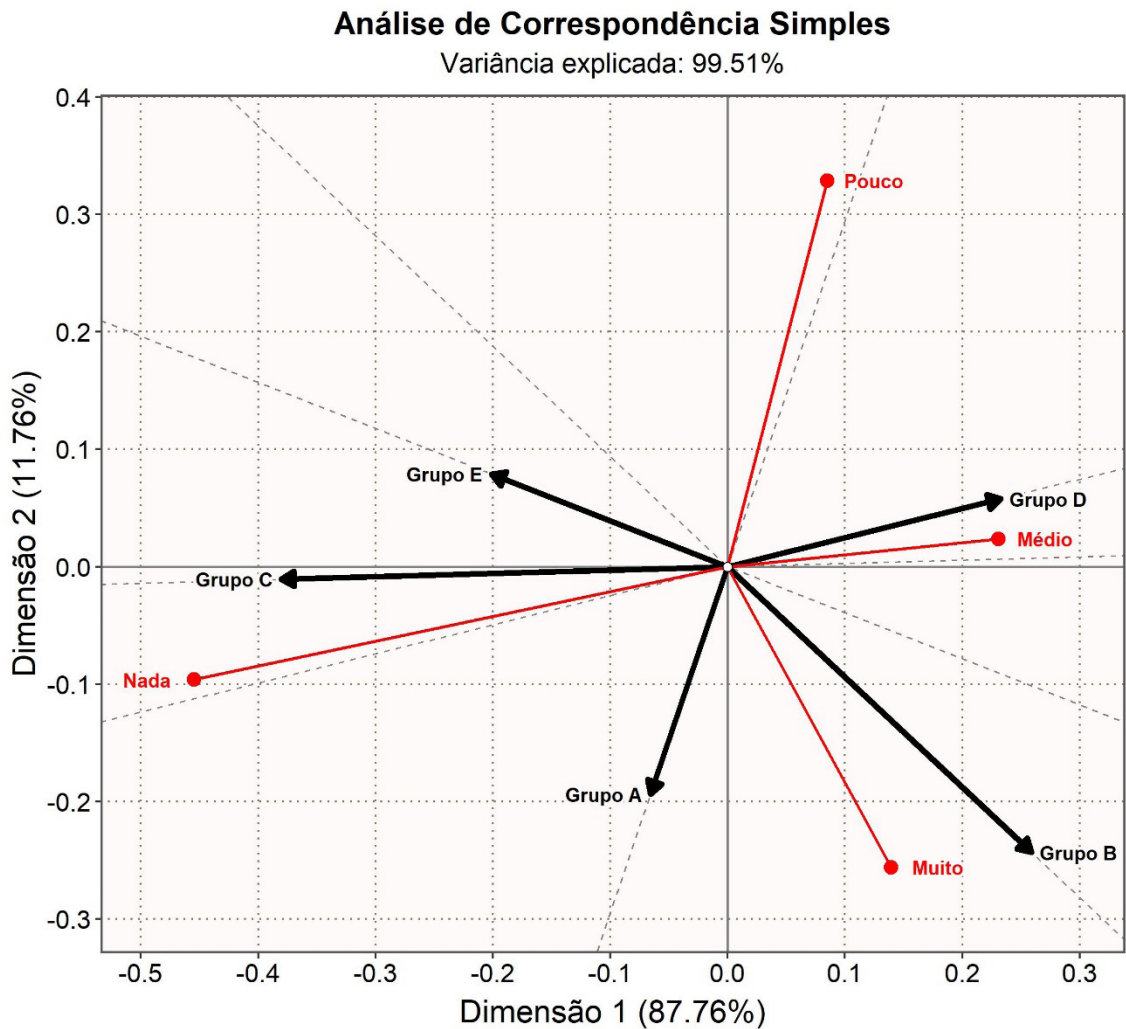


Figura 6. Mapa produzido pela Análise de Correspondência Simples a partir dos dados da Tabela 1

3.1 Uma investigação da área com suporte da AC

O exemplo anterior mostra de que forma uma AC pode auxiliar em evidenciar associações entre variáveis que estão ocultas em tabelas de contingência. Recentemente, Antunes Jr., Ostermann e Cavalcanti (2019) utilizaram a Análise de Correspondência para investigar as relações entre a formação dos orientadores, os referenciais teóricos adotados e as regiões do país onde estão sendo publicadas as dissertações do Mestrado Nacional Profissional em Ensino de Física (MNPEF). A base de dados para a análise foi as dissertações e/ou nos produtos educacionais apresentados no âmbito no MNPEF. Questões como formação do aluno-professor, formação do orientador e coorientador, natureza do produto educacional e articulação do referencial teórico com o produto educacional são alguns dos fatores que foram levados em conta para a compilação dos dados. Como exemplo, será apresentado um exame mais detalhado apenas levando em consideração a formação e histórico de pesquisa do orientador e a articulação do referencial teórico no produto educacional, conforme ilustra a Figura 7.

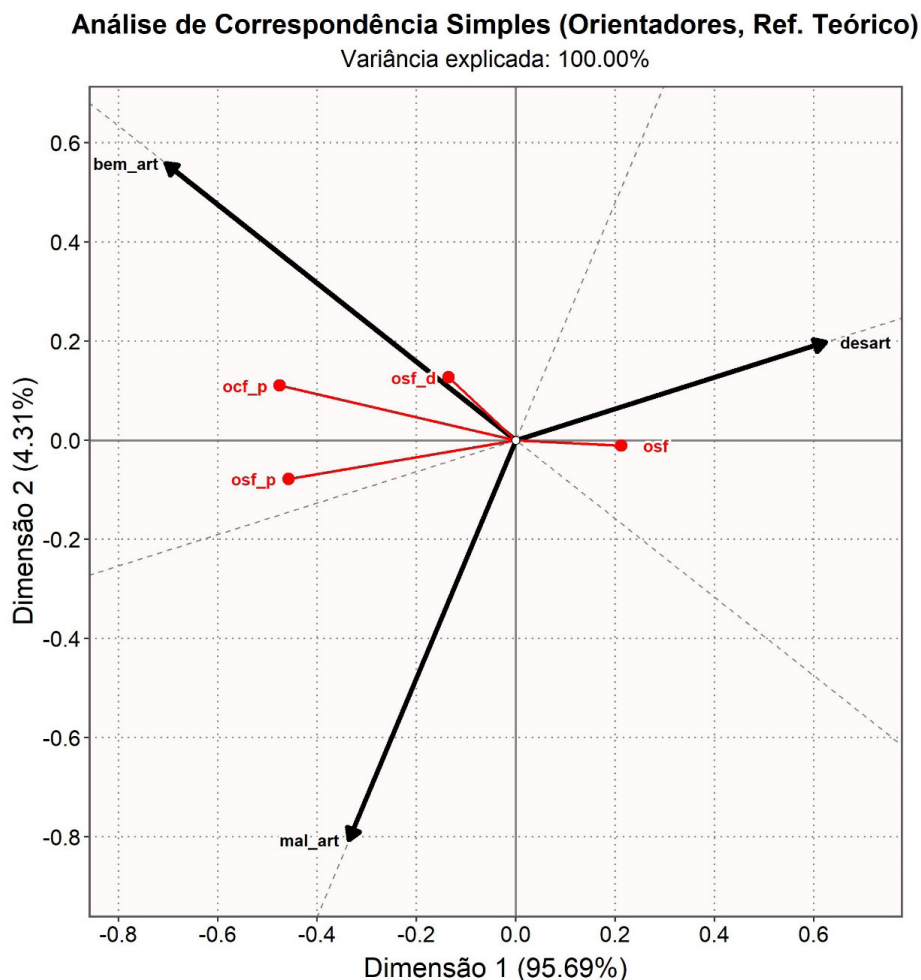


Figura 7. Análise de Correspondência Simples para as variáveis orientadoras e referencial teórico

Fonte: Antunes Jr.; Ostermann; Cavalcanti (2019).

É fácil perceber a partir do mapa que a associação que resulta entre os atributos orientador com formação e produção na área (*ocf_p*) e referencial teórico bem articulado (*bem_art*) é maior do que a associação entre orientador sem formação, mas com produção na área (*osf_p*) e *bem_art*. A visualização é bastante confiável, já que a variância explicada nas duas dimensões é de 100 por cento. É claro também que trabalhos em que o referencial teórico é desarticulado da elaboração do produto educacional (*desart*) são associados bem mais a orientadores sem formação e sem produção na área (*osf*). A partir dessa análise, os autores concluem que uma formação adequada ou larga experiência na área de Ensino por parte dos orientadores resulta em dissertações cujos produtos educacionais articulam-se bem melhor aos referenciais teóricos. Como a visualização da Figura 7 é exata, já que explica 100 por cento da variância, os segmentos curtos são pouco representativos frente aos outros, o que significa que o atributo que diz respeito aos orientadores sem formação e produção em desenvolvimento na área (*osf_d*) não se associa de forma relevante aos níveis de articulação do referencial teórico com o produto educacional. Ainda assim, a categoria *osf_d* tende a se caracterizar mais por referenciais teóricos bem articulados e a categoria *osf* por referenciais teóricos totalmente desarticulados do produto educacional. Essa conclusão se torna mais evidente efetuando essa análise conjunta de múltiplos fatores a partir da AC.

4. Considerações finais

No presente trabalho foram discutidos alguns métodos quantitativos interpretativos a partir de uma abordagem que se distancia das abordagens quantitativas tradicionais. A estratégia foi a de apresentar os métodos a partir da sua efetiva utilização com o apoio de dados obtidos livremente da *internet*. O trabalho tratou especialmente sobre métodos de análise multivariada de dados: o Escalonamento Multidimensional e a Análise de Correspondência. Após a apresentação dos métodos, pesquisas recentes da área da Educação em Ciências foram trazidas para exemplificar a sua efetiva utilização. Todas as visualizações e análises estatísticas foram realizadas com o apoio do ambiente de programação R (R Core Team, 2015).

Em linhas gerais, o Escalonamento Multidimensional (EMD) é um método utilizado para a visualização de dissimilaridades – que podem ser entendidas, para efeitos práticos de compreensão, como distâncias euclidianas – entre variáveis categóricas (Johnson, & Wichern, 2004). Um exemplo de utilização do EMD foi realizado a partir de dados de um questionário aplicado no início dos anos 90 e que infere a percepção de moradores europeus em relação à Ciência e à Tecnologia (Reif; Melich, 1995). Em seguida, com o objetivo de ilustrar o uso do EMD na área da Educação em Ciências, foram sinteticamente demonstrados os resultados de uma pesquisa que investigou as similaridades entre dissertações e referenciais teóricos de um curso de Mestrado Profissional em Ensino de Física (Nascimento, 2016; Nascimento; Ostermann; Cavalcanti; 2017).

Já a Análise de Correspondência (AC) é uma ferramenta que permite a visualização

da associação entre um conjunto de variáveis categóricas em um mapa de dimensões reduzidas (Hair et al., 2009). Neste trabalho, a AC foi utilizada na interpretação da associação entre trabalhadores de uma empresa fictícia e o seu consumo de tabaco (Greenacre, 1984). A exemplificação de utilização da Análise de Correspondência em uma pesquisa recente da área se deu a partir da investigação de Antunes Jr., Ostermann e Cavalcanti (2019) sobre a formação dos orientadores do MNPEF e o uso dos referenciais teóricos no desenvolvimento dos produtos educacionais.

Por fim, vale ressaltar que este texto teve o objetivo de apresentar alternativas possíveis para o uso de métodos quantitativos, na modalidade interpretativa, especialmente para a pesquisa em Educação em Ciências, mostrando que análises dessa natureza podem se afastar da crença de que os dados expressam por si só uma realidade absoluta. Não foi nossa intenção aprofundar questões filosóficas e epistemológicas imbricadas na apropriação dos métodos quantitativos. O peso dessa discussão foi diminuído justamente para se investir na discussão sobre como essas ferramentas podem ser incorporadas no repertório metodológico dos pesquisadores da área. Em um primeiro momento mostrou-se, na prática, como que esses métodos são utilizados para, posteriormente, apresentar as pesquisas recentes na área.

Referências


- Adamuti-Trache, M., & Andres, L. (2008). Embarking on and persisting in scientific fields of study: Cultural capital, gender, and curriculum along the science pipeline. *International journal of science education*, 30(12), 1557–1584.
- Age, L.-J. (2011). Grounded theory methodology: Positivism, hermeneutics, and pragmatism. *The qualitative report*, 16(6), 1599–1615.
- Andres, L., & Adamuti-Trache, M. (2008). University attainment, student loans, and adult life course activities: A fifteen-year portrait of young adults in British Columbia. In R. Finnie, R. E. Mueller, A. Sweetman, & A. Usher (Eds.). *Who goes? Who stays? What matters? Accessing and persisting in post-secondary education in Canada* (pp. 239–275). Retrieved from http://higheredstrategy.com/mesa/pdf/MESA_Andres_Adamuti_Trache.pdf
- Anton, H., & Busby, R. C. (2006). *Álgebra linear contemporânea*. São Paulo, Brasil: Bookman Editora.
- Antunes Jr., E., Ostermann, F., & Cavalcanti, C. J. H. (2019) A subvalorização da formação continuada de professores: dos orientadores à articulação do referencial teórico no contexto do Mestrado Nacional Profissional em Ensino de Física. *Alexandria: Revista de educação em Ciências e Tecnologias*, 12(2), 267–291.
- Babones, S. (2016). Interpretive quantitative methods for the social sciences. *Sociology*, 50(3), 453–469.

- Borg, I., Groenen, P. J., & Mair, P. (2017). *Applied multidimensional scaling and unfolding*: Berlin: Springer. <https://doi.org/10.1007/978-3-642-31848-1>
- Brannen, J. (2017). *Mixing methods: Qualitative and quantitative research*: Routledge.
- Bryman, A. (1984). The debate about quantitative and qualitative research: a question of method or epistemology? *British journal of Sociology*, 75–92.
- Carvalho, R. C. D., Oliveira, I., & Rezende, F. (2009). Tendências da pesquisa na área de educação em ciências: uma análise preliminar da publicação da ABRAPEC. *VII Encontro Nacional de Pesquisa em Educação em Ciências*, 1–12.
- Cerda J., & Valdivia G. (2007). John Snow, la epidemia de cólera y el nacimiento de la epidemiología moderna. *Revista chilena de infectología*, 24(4), 331–334. <https://dx.doi.org/10.4067/S0716-10182007000400014>
- Cho, J. Y., & Lee, E.-H. (2014). Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *The qualitative report*, 19(32), 1–20.
- Cox, T. F., & Cox, M. A. (2000). *Multidimensional scaling*. Nova Iorque, US: Chapman and hall/CRC.
- Cunha, M. V. (2001). John Dewey e o pensamento educacional brasileiro: a centralidade da noção de movimento. *Revista Brasileira de Educação*, 86–99.
- De Leeuw, J., & Mair, P. (2009). Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software*, 31(3), 1–30. <https://doi.org/10.18637/jss.v031.i03>
- De Nooy, W. (2003). Fields and networks: correspondence analysis and social network analysis in the framework of field theory. *Poetics*, 31(5–6), 305–327.
- Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, 14(6), 927–930.
- Ferreira, M. C. (2003). Consumo cultural e espaços sociais: os vestibulandos das universidades públicas na cidade do Rio de Janeiro, 1990. *Opinião Pública*, 9(1), 170–189.
- Greca, I. M. (2002). Discutindo aspectos metodológicos da pesquisa em ensino de ciências: algumas questões para refletir. *Revista Brasileira de Pesquisa em Educação em Ciências*, 2(1), 73–82. Recuperado de <https://periodicos.ufmg.br/index.php/rbpec/article/view/4152>.
- Greca, I. M., & Moreira, M. A. (2001). O uso da análise multidimensional na pesquisa em ensino de ciências. *Revista Brasileira de Pesquisa em Educação em Ciências*, 1(3). Recuperado de <https://periodicos.ufmg.br/index.php/rbpec/article/view/4162>.
- Greenacre, M. (2017). *Correspondence analysis in practice*. Nova Iorque, US: Chapman and Hall/CRC.


- Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. Londres, Inglaterra: Academic Press.
- Greenacre, M. J., & Blasius, J. (1994). *Correspondence analysis in the social sciences: Recent developments and applications*. Londres, Inglaterra: Academic Press.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados*: São Paulo, Brasil: Bookman Editora.
- Hilger, T. R. (2009). *Representações sociais da física quântica* (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul. Porto Alegre, Brasil. Recuperado de: <https://www.lume.ufrgs.br/handle/10183/17743>.
- Hirschfeld, H. O. (1935). *A connection between correlation and contingency*. Trabalho apresentado em Mathematical Proceedings of the Cambridge Philosophical Society.
- Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis*. 6th. New Jersey, US: Pearson Prentice Hall.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, 33(7), 14–26.
- Johnson, W. P., & Myatt, G. J. (2014). *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining 2nd Edition*: John Wiley & Sons.
- Kincheloe, J. L., & Tobin, K. (2009). The much exaggerated death of positivism. *Cultural Studies of Science Education*, 4 (3), 513–528.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.
- Levshina, N. (2015). *How to do linguistics with R: data exploration and statistical analysis*. Amsterdam, Holanda: John Benjamins Publishing Company.
- Medeiros, A. (2002). Metodologia da pesquisa em educação em ciências. *Revista Brasileira de Pesquisa em educação em ciências*, 2(1). 66–72. Recuperado de: <https://periodicos.ufmg.br/index.php/rbpec/article/download/4151/2716>.
- Nascimento, M. M., Cavalcanti, C., & Ostermann, F. (2017). Análise de Correspondência aplicada à pesquisa em Ensino de Ciências. *Enseñanza de las ciencias*, (Extra), 1319-1324. Recuperado de: <https://ddd.uab.cat/record/184468>.
- Nascimento, M. M. (2016). *Análise de produtos educacionais desenvolvidos no âmbito de um mestrado profissional em ensino de física* (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul. Porto Alegre, Brasil. Recuperado de <https://www.lume.ufrgs.br/handle/10183/139386>.
- Nascimento, M. M., Ostermann, F., & Cavalcanti, C. (2017). Análises multidimensional e Bakhtiniana do discurso de trabalhos de conclusão desenvolvidos no âmbito de um mestrado profissional em ensino de Física. *Ciência & educação*. 23(1), 181–196. <https://doi.org/10.1590/1516-73132017001001>.

- Nenadic, O., Greenacre, M. (2007). Correspondence Analysis in R, with two-and three-dimensional graphics: The ca package. *Journal of Statistical Software* 20(3), 1–13. <https://doi.org/10.18637/jss.v020.i03>
- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H., Szoecs, E., & Helene Wagner (2019). *Vegan: Community Ecology Package*. R package version 2.5–4. <https://CRAN.R-project.org/package=vegan>
- Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologist*. Cambridge, Reino Unido: Cambridge University Press.
- R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Recuperado de <https://www.R-project.org>
- Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, *Journal of Statistical Software*, 17(5), 1–25. <https://doi.org/10.18637/jss.v017.i05>
- Reif, K., & Melich, A. (1995). *Euro-barometer 38.1: Consumer protection and perceptions of science and technology, november 1992*: Inter-university Consortium for Political and Social Research.
- Rodrigues, E. V., Borges, A. E., & Pietrocola, M. P. (2018). *História da Ciência no Ensino de Física como representação social dos licenciandos. XVII Encontro de Pesquisa em Ensino de Física*. Campos do Jordão - SP.
- Santos, C. A., & Moreira, M. A. (1991). *Escalonamento multidimensional e análise de agrupamentos hierárquicos*. Porto Alegre, Brasil: Editora da Universidade.
- Santos, C. A. d. (1978). *Aplicação da análise multidimensional e da análise de agrupamentos hierárquicos ao mapeamento cognitivo de conceitos físicos* (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul. Porto Alegre, Brasil. Recuperado de <https://www.lume.ufrgs.br/handle/10183/29922>.
- Schneider, E. M., Fujii, R. A. X., & Corazza, M. J. (2017). Pesquisas quali-quantitativas: contribuições para a pesquisa em ensino de ciências. *Revista Pesquisa Qualitativa*, 5(9), 569–584.
- Shanka, T., Quintal, V., & Taylor, R. (2006). Factors influencing international students' choice of an education destination—A correspondence analysis. *Journal of Marketing for Higher Education*, 15(2), 31–46.
- Souza, A. C., Bastos, R. R., & Vieira, M. d. T. (2010). Análise de Correspondência Simples e Múltipla para Dados Amostrais Complexos. *19º SINAPE: São Pedro/SP*.
- Young, F. W. (2013). *Multidimensional scaling: History, theory, and applications*: Psychology Press.


Matheus Monteiro Nascimento

 <http://orcid.org/0000-0001-8179-5391>
Universidade Federal do Rio Grande do Sul
Departamento de Física
matheus.monteiro@ufrgs.br


Estevão Antunes Júnior

 <http://orcid.org/0000-0002-9989-0335>
Universidade Federal do Rio Grande do Sul
Programa de Pós Graduação em Ensino de Física
eantunesjr@gmail.com

Cláudio José de Holanda Cavalcanti

 <http://orcid.org/0000-0002-2477-3150>
Universidade Federal do Rio Grande do Sul
Departamento de Física
claudio.cavalcanti@ufrgs.br

Fernanda Ostermann

 <http://orcid.org/0000-0002-0594-2174>
Universidade Federal do Rio Grande do Sul
Departamento de Física
fernanda.ostermann@ufrgs.br

Submetido em 04 de julho de 2019

Aceito em 27 de novembro de 2019

Publicado em 26 de dezembro de 2019