Construindo espaços relacionais com a análise de correspondências múltiplas: aplicações nas ciências sociais

Edison Bertoncelo

Brasília | DF
2022

Construindo espaços relacionais com a análise de correspondências múltiplas: aplicações nas ciências sociais

Edison Bertoncelo

Brasília | DF
2022

Fundação Escola Nacional de Administração Pública

Presidente
Diogo Godinho Ramos Costa

Diretora-Executiva
Rebeca Loureiro de Brito

Diretora de Altos Estudos
Diana Coutinho

Diretor de Educação Executiva
Rodrigo Torres

Diretor de Desenvolvimento Profissional
Paulo Marques

Diretora de Inovação
Bruna Silva dos Santos

Diretora de Gestão Interna
Alana Regina Biagi Silva Lisboa

Revisão gramatical
Renata Mourão

Projeto gráfico e editoração eletrônica
Amanda Soares

Catalogado na fonte pela Biblioteca Graciliano Ramos da Enap

B547c Bertoncelo, Edison

Construindo espaços relacionais com a análise de correspondências múltiplas: aplicações nas ciências sociais / Edison Bertoncelo. -- Brasília: Enap, 2022.
143 p. : il.

Inclui bibliografia.
ISBN: 978-65-87791-12-8

1. Ciências Sociais Aplicadas. 2. Administração Pública. 3. Inovação. 4. Análises Múltiplas. 5. Pesquisa. 6. Análise de conteúdo. I. Título.

CDU 303.732.3

Bibliotecária: Tatiane de Oliveira Dias – CRB1/2230

Escola Nacional de Administração Pública
SAIS – Área 2-A – 70610-900 — Brasília-DF, Brasil

Sumário

Lista de tabelas

Lista de figuras

Introdução

A análise de correspondências múltiplas e a estatística multivariada

O objetivo principal deste livro é explorar as possibilidades da aplicação de uma técnica denominada **análise de correspondências múltiplas** (doravante ACM) nas ciências sociais.[1] A exposição dos conteúdos seguirá uma intenção prática, orientada para auxiliar o leitor na utilização dessa técnica.[2] Pretende-se que a leitura do material possibilite à leitora e ao leitor interpretar criticamente os resultados produzidos por meio dessa técnica (e outras similares) e utilizá-la, de modo criativo e rigoroso, em suas pesquisas.

[1] Considero a análise de correspondências múltiplas uma técnica, no sentido de uma ferramenta ou instrumento, dentro de um conjunto mais amplo de métodos estatísticos multivariados.

[2] A discussão das abordagens da geometria que dão fundamento a essa técnica será feita apenas com o intuito de tornar possível a interpretação dos resultados. Para um aprofundamento nessa discussão, ver Le Roux e Rouanet (2004); Greenacre e Blasius (2006); e Blasius e Greenacre (2014).

A ACM é uma técnica de análise de dados utilizada para "descrever, explorar, sumarizar e visualizar informações contidas em uma tabela de dados de *N* indivíduos descritos por *Q* variáveis categóricas" (Husson; Josse; 2014, p. 165). Uma base de dados desse tipo é geralmente obtida a partir da aplicação de **questionários estruturados** a uma amostra de uma população qualquer, mas também pode ser o resultado da codificação dos atributos de casos estudados a partir de uma técnica qualitativa (por exemplo, por meio de entrevistas ou de uma pesquisa documental).

A ACM faz parte de um conjunto mais amplo de técnicas denominado **análise geométrica de dados** (doravante AGD).[3] O objetivo das técnicas reunidas sob esse rótulo reside na **construção de nuvens de pontos**, quer dizer, "um conjunto finito de pontos em um espaço geométrico" (Le Roux; Rouanet, 2010, p. 14). Na AGD,

> [...] as linhas ou colunas de uma matriz de dados são entendidas como pontos em um espaço euclidiano de elevada dimensão, e o método busca redefinir as dimensões do espaço de modo que as principais dimensões capturem a maior variância possível, possibilitando descrições dos dados em menor dimensão (Greenacre; Blasius, 2006, p. 5).

Comum a todas as técnicas incluídas sob esse rótulo, está o uso de ferramentas visuais (as chamadas **nuvens de pontos**) como uma estratégia de análise exploratória que ganhou força a partir dos anos de 1960, concomitantemente ao desenvolvimento de aparatos computacionais (Lebart; Saporta, 2014).

Sob a denominação de AGD, são incluídas três técnicas que se diferenciam em termos do tipo de variável e/ou das características da matriz de dados. A **análise de correspondências simples** (ACS) é usada para analisar tabelas de contingência formadas por duas variáveis categóricas (por exemplo, a percepção do tipo de sociedade em que se vive e o país do

[3] O uso desse termo foi sugerido pelo filósofo e matemático Patrick Suppes (Universidade de Stanford) em 1996 (Le Roux, 2014).

respondente)[4]. Para matrizes de dados com o formato **indivíduo x variável** (ou seja, em que cada linha corresponde a um "indivíduo" e cada coluna, a uma variável), as técnicas a serem usadas são: a **análise de componentes principais (ACP)**[5], quando se analisam variáveis **quantitativas** (ou seja, variáveis que possem valores numéricos, como o número de livros que as pessoas leram em um ano); e a **análise de correspondências múltiplas**, quando se trata de variáveis **categóricas**.

A distinção entre variáveis quantitativas e categóricas é pertinente tecnicamente (quer dizer, para a escolha de qual técnica usar)[6], mas não é essencial do ponto de vista metodológico.[7] As variáveis categóricas são aquelas que possuem um número finito de categorias, entre as quais pode haver algum ordenamento (**ordinal**) ou não (**nominal**).[8] Em outras palavras, uma varável é categórica quando "a escala de mensuração é um conjunto de categorias" (Agresti; Finlay, 2012, p. 28). Sua escala de mensuração pode ser **nominal** (por exemplo: local de moradia ou raça/cor) ou **ordinal** (por exemplo: frequência de práticas culturais, gosto musical mensurado em termos de uma escala que vai de zero a cinco). Uma variável ordinal pode ser o resultado da recodificação de uma variável quantitativa em um número finito de categorias (por exemplo: a variável anos de estudo completo pode ser recodificada em faixas de escolaridade).

A ACM é entendida geralmente como uma **variante** e **extensão da ACS**: enquanto esta é uma técnica para analisar tabelas de contingência pela visualização conjunta de linhas e colunas em um espaço de baixa dimensão, a ACM parte de uma "tabela disjuntiva binária", ou matriz indicadora, podendo representar indivíduos e modalidades (Lebart; Saporta, 2014). Os indivíduos podem ser os respondentes de um questionário, organizações,

[4] Cf. Hjellbrekke (2019, p. 3).

[5] Para um exemplo do uso da ACP em uma pesquisa sociológica, ver Peroza, Lebaron e Leite (2015).

[6] A mesma discussão pode ser feita em relação aos usos dos diferentes tipos de regressão, de testes de correlação ou de associação etc. (Schroeder; Sjoquist; Stephan, 1986).

[7] Isso porque o tipo de variável (e mesmo o método de análise) não deriva inexoravelmente da natureza do dado, mas das escolhas metodológicas efetuadas em função das questões a que se quer responder com a pesquisa (Pires, 2008; Cano, 2012).

[8] Veremos que a diferenciação entre variáveis categóricas nominais ou ordinais é pertinente para o uso correto da ACM.

grupos, sociedades ou quaisquer unidades de análise. As modalidades são as informações que reunimos acerca dos indivíduos.

Uma matriz de dados do tipo indivíduo x variável é tipicamente resultante da aplicação de questionários estruturados a uma amostra (estatisticamente representativa ou não) de um universo social qualquer: para cada questão q, há um conjunto J_q de categorias de respostas (também designadas por **modalidades**) e cada indivíduo i escolhe uma e apenas uma única categoria no conjunto J_q.[9] Nos casos da ACP e, sobretudo, da ACM,

> [A] instância típica [de aplicação] é a análise de questionários, quando o conjunto de questões é suficientemente amplo e, ao mesmo tempo, diversificado o bastante para cobrir vários temas de interesses (entre os quais, algum equilíbrio é alcançado), de forma a produzir representações multidimensionais significativas (Rouanet, 2006, p. 142).

A AGD, como uma abordagem estatística, foi inicialmente elaborada nos trabalhos do matemático francês Jean-Paul Benzécri nos anos de 1970.[10] Seus trabalhos foram considerados importantes para contrabalançar a atenção excessiva que se dava, então, aos modelos confirmatórios na estatística (Greenacre; Blasius, 2006). Nas palavras de Benzécri (1973 *apud* Greenacre; Blasius, 2006, p. 6), "o modelo deve seguir os dados, e não o contrário".[11] A partir da década de 1980, a AGD passou a figurar mais frequentemente em publicações de língua inglesa e em manuais de metodologia.[12] Houve também um notável crescimento no número de publicações sobre essa técnica em periódicos científicos (Greenacre;

[9] Ao longo desta exposição, utilizaremos os termos categorias ou modalidades indistintamente.

[10] O livro de referência é *Analyse de donnés*, publicado em 1973. Para uma história da ACS e ACM, ver Lebart e Saporta (2014).

[11] "The model should follow the data, not the reverse."

[12] Com ênfase nos trabalhos de Michael Greenacre.

Blasius, 2006, p. 1). Nas ciências sociais brasileiras, são cada vez mais numerosos os trabalhos que empregam as diferentes técnicas da AGD.[13]

As três ideias básicas da AGD são:

> i. **modelos geométricos**: construção de nuvens de pontos em um espaço geométrico. Veremos que a ACM produz dois resultados principais: a **nuvem de modalidades** e a **nuvem de indivíduos**;
>
> ii. abordagem formal, baseada na álgebra linear;
>
> iii. "descrição em primeiro lugar": análise descritiva precede a modelagem probabilística. "Os resultados básicos dos métodos geométricos são **estatísticas descritivas**, no sentido técnico de que não dependem do tamanho da base de dados" (Le Roux; Rouanet, 2010, p. 14).

Embora não exista uma relação de correspondência direta entre teoria e método (Platt, 1986), há uma espécie de "afinidade eletiva" entre a teoria sociológica de Pierre Bourdieu e o uso da AGD. Conforme argumenta Bourdieu:

> Eu uso muito Análise de Correspondências, porque penso que é essencialmente um procedimento relacional cuja filosofia expressa plenamente o que, a meu ver, constitui a realidade social. É um procedimento que "pensa" em relações, assim como tento fazer com o conceito de campo (Bourdieu, 1991 *apud* Lebaron, 2009, p. 13).

Essa lógica **relacional** implica que as práticas sociais não têm significado em si mesmas, mas apenas na estrutura de contrastes ou afinidades entre elas. A AGD reúne técnicas que possibilitam operacionalizar essa concepção relacional do social.

São muitas as **técnicas estatísticas multivariadas** – ou seja, as que envolvem diversas variáveis, sejam elas quantitativas e/ou categóricas – empregadas nas ciências sociais: além da AGD, mencionemos os diversos

[13] Cf. Klüger (2019). Além do artigo referido, há também um *paper* publicado nos anais do congresso da Anpocs sobre os usos da ACM nas Ciências Sociais (Bertoncelo, 2016).

tipos de regressão (linear, logística, multinomial), análise fatorial, análise de classes latentes, os modelos log-lineares etc. A escolha pela(s) técnica(s) mais adequada(s) – ou de uma combinação entre elas – dependerá das questões e dos objetivos da pesquisa, da natureza dos dados e do desenho amostral (Silva, 2018).

As técnicas estatísticas **não** são em si mesmas exploratórias, explicativas ou confirmatórias. Tais qualificativos têm a ver com interesses de pesquisa, e não com propriedades supostamente inerentes a esta ou aquela técnica (Le Roux; Rouanet, 2004, p. 19-20). Não me parece correto, portanto, definir a AGD como um conjunto de técnicas essencialmente **exploratórias** em oposição a outras (como a regressão), que seriam, "por natureza", **confirmatórias** e/ou **explicativas**. De fato, como veremos nos próximos capítulos, um elemento importante da AGD é a análise de dados estruturados [*structured data analysis*], que nos auxilia a detectar os fatores determinantes dos padrões empiricamente evidenciados ou, mesmo, a prever determinados aspectos das práticas dos agentes com base no conhecimento de algumas de suas propriedades (relacionais).

Ao mesmo tempo, é possível identificar algumas oposições nos usos da estatística multivariada nas ciências sociais, a saber:

i. Em contraste com o uso convencional da estatística multivariada baseado em "indicadores numéricos" (como os coeficientes de regressão) e no "sistema de asteriscos" (que indicam o nível de significância estatística), na AGD, a construção de **nuvens de pontos** (de modalidades e de indivíduos) é o elemento central. Trata-se, conforme sustentam Le Roux e Rouanet (2004), de duas "concepções distintas do papel da estatística": de um lado, uma "sociologia das variáveis" e, de outro, a ênfase **na construção de espaços relacionais** (na sociologia bourdieusiana, o espaço social ou das classes sociais e o espaço simbólico, dos estilos de vida ou das tomadas de posição).

ii. Diferentemente dos usos da estatística multivariada orientados para a construção e teste de modelos (exemplos: análise de classes latentes e modelos log-lineares) – ou seja, em que, após a especificação de um número qualquer de modelos, é escolhido, com base nos critérios disponíveis, aquele que mais se "ajusta" aos dados, considerando o equilíbrio entre exaustividade e parcimônia –, os

usos da AGD geralmente privilegiam uma **abordagem indutiva**, em que "os modelos devem seguir os dados, e não o contrário" (Benzécri, 1973 *apud* Greenacre; Blasius, 2006, p. 6). É a partir da observação das posições relativas das modalidades e indivíduos nas nuvens de pontos (daí, espaços relacionais) que são apreendidas as principais dimensões que estruturam as afinidades e contrastes entre práticas dos agentes e suas propriedades.

iii. Por fim, os usos da AGD estão assentados em uma concepção de causalidade que difere daquilo que Andrew Sayer (1992) denominou de **análise causal padrão**, que consiste em isolar o efeito de cada variável independente sobre a variável dependente em uma abordagem quase experimental. Diferentemente, na AGD, o que importa é **a estrutura de relações** entre um conjunto de variáveis e suas categorias (Hjellbrekke, 2019), preocupação que está assentada em uma concepção de causalidade que poderíamos denominar de "estrutural": a causalidade social consiste nos "efeitos globais de uma estrutura complexa de inter-relações, que não se reduz à combinação dos múltiplos 'efeitos puros' das variáveis independentes" (Lebaron, 2009, p. 12). A ACM (e, mais amplamente, a AGD), portanto, não pressupõe uma diferenciação entre variáveis dependentes e independentes ou a concepção de causalidade linear e aditiva em que se baseia. Em vez disso, o uso da ACM se presta muito bem ao objetivo de apreender o "sistema completo de relações que constituem o verdadeiro princípio da força e da forma específicas dos efeitos registrados em determinada correlação particular" (Bourdieu, 2008, p. 98).

Um exemplo: como os gostos culturais se relacionam entre si?

A matriz de dados que será utilizada nos exemplos a seguir pode ser visualizada na Tabela 0.1. Como vimos, a matriz de dados na ACM tem o formato de uma tabela **indivíduos por variáveis**. Cada linha representa um indivíduo; cada coluna, uma variável. Em cada célula, há uma modalidade que caracteriza o indivíduo. No caso dessa matriz, cada célula traz a resposta do indivíduo a uma determinada questão. Por exemplo: você gosta de *rock*? As categorias de resposta possíveis incluem "sim", "não" ou "não sabe".

Tabela 0.1 | Matriz de dados sobre práticas e gostos culturais

	Você gosta de...?							
Indivíduo	**Música romântica**	***Rock***	**Clássica**	**Sertanejo**	***Rap***	***Funk***	**MPB**	**Samba**
1	sim	sim	sim	sim	não	não	sim	sim
2	sim	sim	sim	sim	sim	sim	sim	sim
3	sim	não	não	sim	não	não	não	não
4	não	não	não	sim	não	não	não	não
5	sim	não	não	sim	não	não	sim	não
6	não	sim	não	não	sim	não	não	não
7	sim	não	não	sim	não	não	não	não
8	não	não	não	sim	não	não	sim	não
9	sim	não	não	sim	não	não	sim	sim
10	sim	não	não	sim	não	não	não	não
11	sim	não	sim	sim	não	não	sim	sim
12	sim	não	não	sim	não	não	sim	não

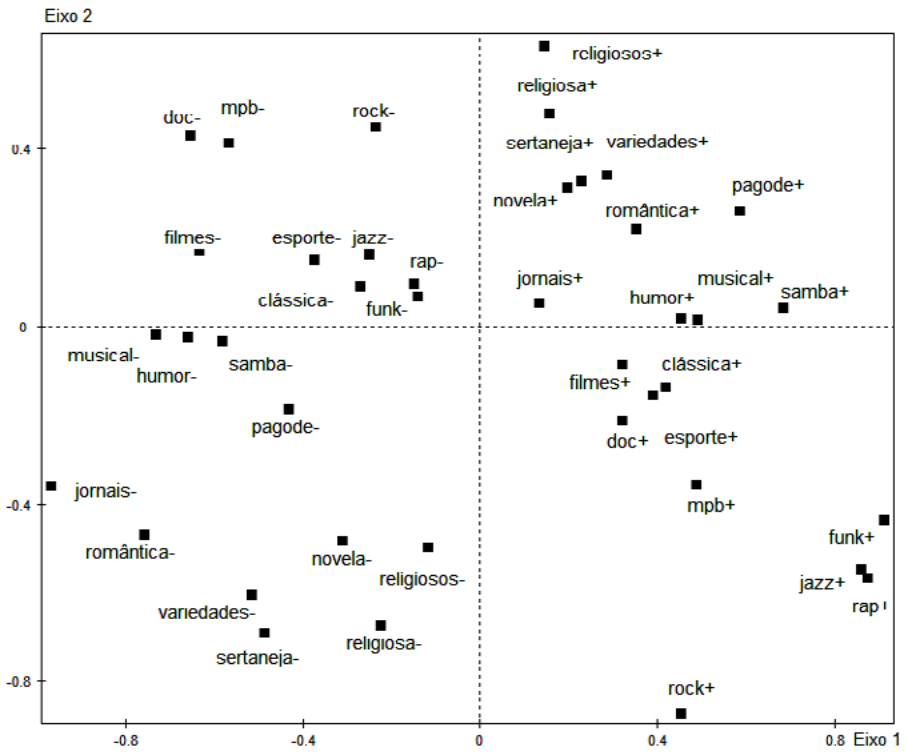
Fonte: elaboração própria.

Consideremos a temática do gosto cultural, mais especificamente as relações entre as preferências em termos de gêneros musicais e de programas de televisão. Aqueles que gostam de *rock* gostam tipicamente de quais outros gêneros musicais? E quais mais comumente rejeitam? E em termos de preferências por gêneros televisivos: o gosto pelo *rock* está associado ao gosto por documentários, programas religiosos, jornais? Ou, diferentemente, o gosto pelo *rock* está associado à rejeição por novelas? Pensando agora na distribuição das possíveis combinações de

preferências entre os agentes, podemos perguntar: a possível combinação das preferências por *rock*, MPB, documentários e esportes é mais comum entre homens ou mulheres, jovens ou idosos, entre indivíduos mais ou menos escolarizados? Em suma, sabendo o conjunto de propriedades sociais que caracterizam um agente, qual é (quais são) o(s) repertório(s) de gostos mais provável(is)?

A resposta a tais questões exige um tipo de **análise relacional** que não está orientada predominantemente para a mensuração do peso de certas variáveis na explicação da probabilidade de um indivíduo gostar de determinado bem cultural (ou seja, não se quer saber qual é o efeito da variável escolaridade, controlada por outros fatores, na probabilidade de o indivíduo gostar em comparação a não gostar de *rock*). Em vez disso, a ênfase recai na estrutura de relações entre as modalidades das variáveis (no caso, entre o gosto ou a rejeição por gêneros culturais): ir a óperas em relação a ir a shows de *rock*, ou visitar museus, frequentar (certos) restaurantes, praticar (determinados) esportes etc. O uso da ACM possibilita, assim, minimizar os riscos de uma **leitura substancialista** das relações das práticas entre si e delas com as propriedades sociais dos agentes (como, por exemplo, considerar que o gosto pela música erudita seja sempre um indicador preciso de um pertencimento de classe específico). Por meio da observação das proximidades ou distâncias relativas entre as modalidades e os indivíduos, é possível reconstruir **indutivamente** os principais contrastes e afinidades entre as práticas sociais e seus agentes.

Figura 0.1 | Nuvem de modalidades projetadas no plano fatorial formado pelos eixos 1 e 2



Fonte: elaboração própria.

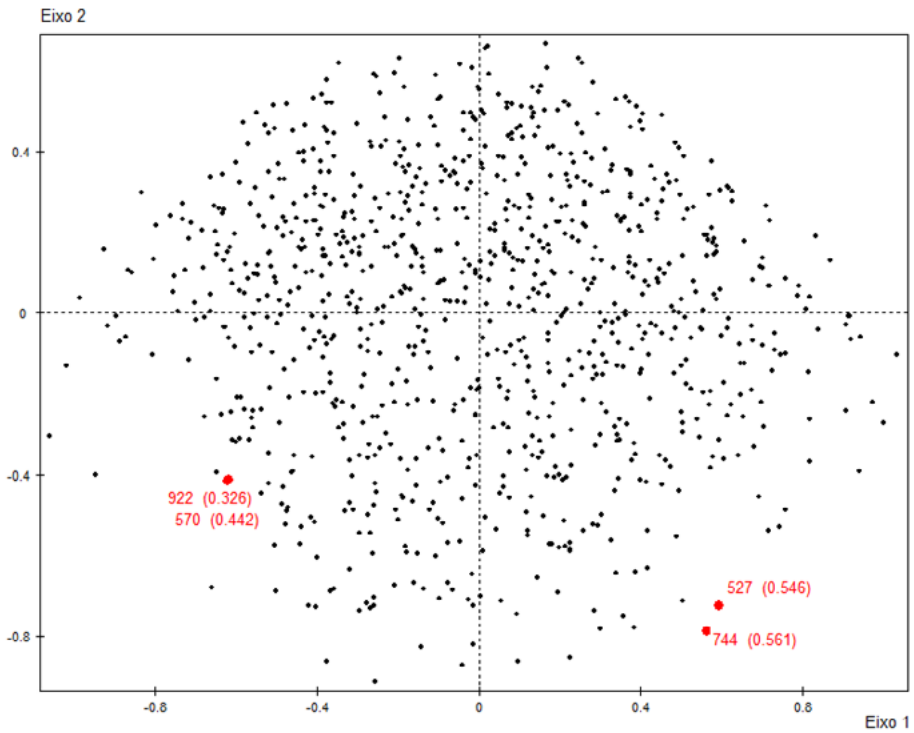
Na Figura 0.1, vemos uma nuvem de modalidades construída pelo cruzamento dos eixos 1 e 2. Tais modalidades indicam o gosto por ou a rejeição a diferentes gêneros musicais e programas de televisão. Como uma análise preliminar, notemos que quanto mais relativamente próximos entre si estiverem os pontos de duas modalidades no plano, mais frequentemente elas tendem a ser escolhidas pelas mesmas pessoas. Ou seja, a proximidade entre duas categorias pode ser interpretada como proximidade entre grupos de indivíduos. No limite, se todas as pessoas que gostam de novela também gostassem de programas religiosos, os pontos que representam essas duas modalidades apareceriam no

mesmo lugar. Diferentemente, se ninguém que gosta de novelas gostasse de programas religiosos, tais pontos estariam em localizações opostas no plano fatorial.

Observando a Figura 0.1, vemos que há um primeiro contraste importante: à esquerda do eixo horizontal, estão todas as categorias que indicam rejeição e, à direita, as modalidades que indicam, ao contrário, o gosto por diferentes gêneros musicais e por programas de televisão. Em relação ao eixo vertical, vemos contrastes (e combinações) mais sutis: acima, estão as modalidades que indicam preferências por programas e músicas religiosas, por programas de variedades, novelas, música sertaneja e romântica, indicando que tendem a ser escolhidas pelas mesmas pessoas. Em contraste, na região inferior do eixo vertical, estão as modalidades que indicam preferências pelo *rock* e MPB, por programas de esportes e documentários, entre outros.

Na Figura 0.2, temos a nuvem de indivíduos construída ao redor dos mesmos eixos. Os pontos representam os indivíduos da amostra, cujas posições relativas dependem das respostas dadas às questões sobre gosto. Quanto mais próximos entre si estão os pontos, maior a similaridade dos padrões de respostas dos indivíduos; diferentemente, quanto mais distantes, maior a dissimilaridade entre eles.

Figura 0.2 | Nuvem de indivíduos projetados no plano fatorial formado pelos eixos 1 e 2



Fonte: elaboração própria.

Notemos as posições relativas dos indivíduos “570” e “922” na porção inferior à esquerda do eixo horizontal. Por serem as nuvens de modalidades e de indivíduos perfeitamente simétricas, sabe-se que os pontos que representam tais indivíduos estão localizados nessa região pelo fato de eles rejeitarem a maior parte dos gêneros musicais e de programas de TV. Ademais, estando representados por pontos bem próximos entre si (embora não exatamente ocupando o mesmo lugar), possuem padrões de respostas bastante similares no que se refere às questões consideradas. De fato, o primeiro indivíduo gosta de apenas um gênero musical (MPB) entre os onze mencionados no questionário e três tipos de programas televisivos (variedades, filmes e documentários) entre os nove mencionados. Já o

segundo indivíduo gosta de apenas dois gêneros musicais (*rap* e pagode) e de dois programas de TV (esportes e filmes). Nos casos, os padrões de respostas se assemelham por incluírem muitas respostas que indicam rejeição aos gêneros culturais presentes no questionário.

Do outro lado do eixo horizontal, também em sua porção inferior, estão os pontos dos indivíduos "527" e "744". Por estarem relativamente próximos entre si, sabemos que tais indivíduos têm padrões de respostas similares e, estando na referida porção do eixo (dada a perfeita simetria entre as nuvens), sabemos também que mencionaram gostar de mais gêneros culturais e, sobretudo, daqueles gêneros cujas modalidades também estão aí localizadas: de fato, o indivíduo "527" rejeita apenas dois gêneros musicais (pagode e religiosa) e gosta de vários programas televisivos, como filmes, jornais, documentários, humor e musicais (rejeita, entre outros, novela, variedades e programas religiosos). O indivíduo "744", de forma similar, aprecia diversos gêneros musicais, com exceção de sertanejo, pagode e música religiosa, e também diversos tipos de programas televisivos, excetuando-se as novelas, os programas de humor e os musicais. Vemos, assim, que os padrões de respostas dos dois indivíduos são não apenas muito similares entre si, o que explica a proximidade relativa dos pontos que os representam no plano formado pelos dois eixos, como também são bem distintos dos padrões de respostas dos indivíduos antes mencionados, o que, por sua vez, explica a elevada distância relativa entre eles.

Em suma, a ACM revela-se uma técnica extremamente útil quando a pesquisa tem por objetivo construir espaços relacionais, representando, sob a forma de nuvens de pontos, as distâncias/proximidades relativas de conjuntos de indivíduos e suas práticas, interesses, opiniões e/ou disposições, para, daí, apreender as principais oposições e afinidades entre eles ou elas e os fatores que as estruturam.

Plano do livro

Este livro é composto por cinco capítulos, além desta introdução. Nos dois próximos capítulos, abordarei os principais conceitos, procedimentos e variantes da ACM, como a construção das nuvens de modalidade e de indivíduos, e a análise de dados estruturados, que consiste em examinar os fatores que estruturam os espaços relacionais. No capítulo 3, tratarei de uma variante da ACM, a **análise de classes específicas**, que possibilita analisar a estrutura de subgrupos considerando sua posição relativa na nuvem global de indivíduos. No capítulo 4, serão exploradas as possibilidades heurísticas da combinação entre a ACM e a **classificação hierárquica ascendente**, que é uma técnica específica de construção de agrupamentos (*clusters*). No capítulo 5, que também servirá como conclusão, tratarei das dificuldades mais comuns relacionadas com o uso da ACM para a construção de espaços relacionais a partir da utilização de fontes secundárias.

Capítulo 1

Análise de correspondências múltiplas: construindo e interpretando a nuvem de modalidades

Para que serve a ACM?

Examinar simultaneamente as nuvens de indivíduos e de modalidades é a principal possibilidade oferecida pela ACM para a pesquisa empírica, fazendo dela uma ferramenta ao mesmo tempo **exploratória** e **preditiva**. Exploratória porque permite representar visualmente as relações entre as modalidades de variáveis categóricas por meio de proximidades e distâncias relativas em um espaço cartesiano de duas dimensões. Preditiva porque, uma vez reconstruída a estrutura de afinidades e contrastes

entre as modalidades, é possível projetar os indivíduos nesse plano bidimensional, examinar como suas propriedades (por exemplo, capital cultural, capital econômico, classe social, idade, gênero etc.) estão correlacionadas com tal estrutura e também os padrões de dispersão dos indivíduos, caracterizados em função de uma ou mais propriedades, no referido plano (Husson; Josse, 2014).

Tais possibilidades nos dão a oportunidade de investigar velhas (e novas) problemáticas a partir de outros ângulos. Daí, a questão: para que serve a ACM? Considerando que qualquer técnica ou método deve ser escolhido em função das questões de pesquisa, quais são as problemáticas para as quais essa técnica ou método pode nos ajudar a encontrar soluções? No capítulo anterior, argumentei que a ACM – e também a ACS – foi utilizada por Pierre Bourdieu em suas pesquisas por possibilitar a operacionalização da noção de campo entendido como "'sistema' ou 'espaço' estruturado de posições ocupadas pelos diferentes agentes" (Lahire, 2017, p. 65). Só podemos compreender as práticas e as estratégias dos agentes se as relacionarmos às posições ocupadas pelos agentes no campo, ou melhor, ao conjunto de relações entre as posições do campo. Quer dizer, **o sentido e o valor das práticas e das propriedades dos agentes só podem ser apreendidos por meio da reconstrução dessa estrutura de relações**. O valor de um diploma de ensino superior vai variar bastante em função da raridade relativa dessa propriedade; assim como o valor simbólico atribuído ao *rock* também variará conforme as relações entre esse gênero e outros gêneros culturais e conforme as propriedades sociais típicas de seus consumidores.

Tomemos, como exemplo, o campo das classes sociais da sociedade francesa investigada por Bourdieu em suas pesquisas ao longo das décadas de 1960 e 1970. Esse campo é, na verdade, o macrocosmo representado pelo espaço social global (circunscrito ao território nacional). O espaço social é definido como "o sistema formado pelo conjunto das posições ocupadas pelos agentes em uma dada formação social" (Nogueira, 2017, p. 177). As posições relativas dos agentes no espaço social são estruturadas pela distribuição das propriedades efetivas em um dado universo social que, no caso da sociedade francesa de então, correspondiam a diferentes tipos de capital, sobretudo o

capital econômico e o capital cultural. As posições na distribuição dessas propriedades são diferenciadas em três dimensões: o volume global do capital, a composição ou estrutura do capital (maior ou menor peso das diferentes espécies de capital) e a modalidade de apropriação do capital (que se refere às trajetórias dos agentes pelo espaço social: acumulação de capital, reconversão entre capitais etc.). Tendo diversos indicadores desses diferentes tipos de capital e de origem social, Bourdieu pôde reconstruir o espaço das posições sociais, ao qual se sobrepõe o espaço dos estilos de vida, utilizando a ACM (Bourdieu, 2008, p. 118-119).

O exemplo anterior, baseado n'*A distinção* (Bourdieu, 2008), evidencia que a ACM é uma técnica muito útil para operacionalizar a noção teórica de campo e, mais precisamente, no estudo dos processos de diferenciação e hierarquização. A partir dos estudos de Bourdieu, os usos da ACM nas ciências sociais, sobretudo na sociologia, têm mostrado que a técnica pode nos ajudar a apreender o sentido relacional das propriedades dos agentes (suas práticas, disposição ou recursos) e a reconstruir a formação de fronteiras sociais e simbólicas indutivamente. No caso dos estudos de classe, por exemplo, podemos partir do exame empírico da distribuição dos diferentes tipos de capital entre os agentes para apreender a multidimensionalidade do espaço assim construído (ao invés de delinear as fronteiras de classe *a priori*). Ou, então, num estudo sobre estilos de vida ou sobre opiniões políticas, podemos observar como diferentes práticas, gostos ou preferências se distribuem socialmente e se combinam tipicamente, sem assumir de antemão que tal ou qual prática, gosto ou tomada de posição é mais ou menos legítima ou, então, indicativa de tal ou qual pertencimento social.

Matrizes de dados: a ACM como uma extensão da ACS

A ACS é aplicada a uma tabela de contingência, resultante do cruzamento entre duas variáveis categóricas, em que cada célula representa o número de observações de uma dada combinação entre as categorias, como se vê na Tabela 1.1 a seguir.

Tabela 1.1 | Cruzamento das variáveis "leitura de livros" e "audiência à TV"

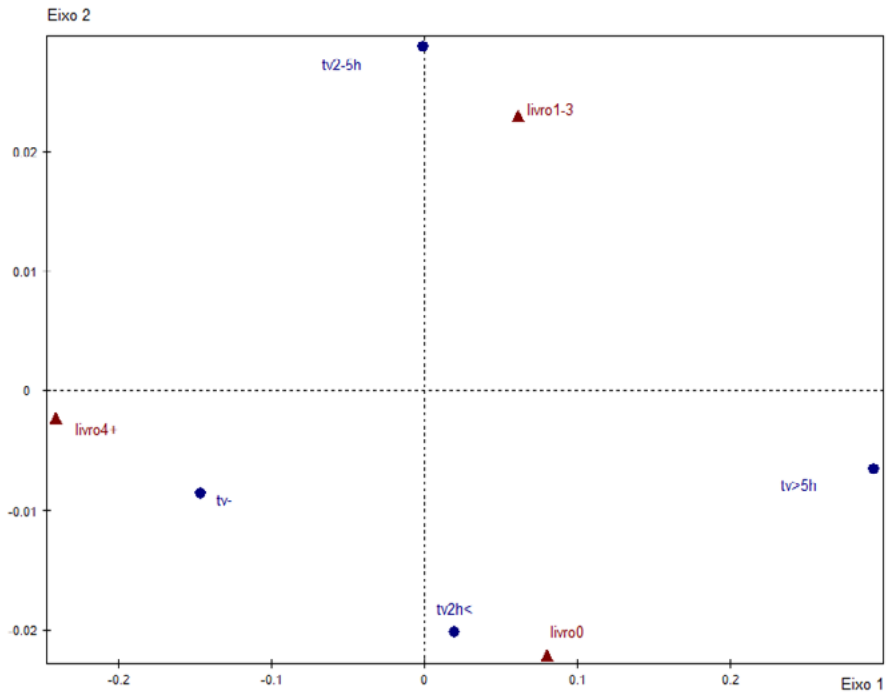
	Nenhum (livro0)	1 a 3 (livro1-3)	4 mais (livro4+)	Total
Mais de 5 horas (tv>5h)	56	54	13	**123**
% linha	45,4	44,16	10,44	**100**
% coluna	15	14	6	**13**
Entre 2 e 5 horas (tv2-5h)	111	121	68	**300**
% linha	36,98	40,34	22,69	**100**
% coluna	30	32	31	**31**
Menos de 2 horas (tv2h<)	107	104	59	**270**
% linha	39,56	38,5	21,94	**100**
% coluna	29	27	27	**28**
Raramente ou nunca assiste (tv-)	99	101	81	**281**
% linha	35,16	35,97	28,87	**100**
% coluna	27	27	37	**29**
Total	**372**	**381**	**221**	**974**

Fonte: elaboração própria.

Percebemos algumas associações importantes entre as categorias das variáveis: entre a leitura de 4 livros ou mais, de um lado, e a ausência ou quase ausência de audiência à TV, de outro; ou, então, a maior audiência à TV e a pouca leitura de livros. A utilização da ACS possibilita representar

visualmente as associações entre as categorias num plano formado por dois eixos. É possível observar, na Figura 1.1, as proximidades relativas entre a leitura de um a três livros e assistir à TV entre duas a cinco horas por dia, cujas modalidades estão localizadas na região superior do eixo vertical; entre ler 4 livros ou mais e assistir à TV raramente ou nunca, na região inferior à esquerda; entre a ausência da prática de leitura de livros e assistir à TV por menos de duas horas por dia, na região inferior do eixo vertical e à direita do eixo horizontal e, por fim, assistir à TV por mais de cinco horas bem à direita do eixo horizontal, na região oposta em que está localizada a modalidade que indica a leitura de quatro livros ou mais e no mesmo quadrante daquela que indica a ausência de leitura de livros.

Figura 1.1 | Nuvem de modalidades resultante da ACS aplicada à tabela formada pelas variáveis "leitura de livros" e "audiência à TV", projetadas no plano cartesiano formado pelos eixos 1 e 2



Fonte: elaboração própria.

A ACM pode ser entendida como uma extensão da área de aplicação da ACS, não se restringindo a uma tabela de contingência. Como é possível estender o escopo de aplicação para examinar as relações entre as categorias de mais de duas variáveis? Tal extensão é possibilitada pela utilização da chamada **matriz indicadora binária** (Tabela 1.2). Uma matriz desse tipo possui valores de zero e um. Os indivíduos são listados nas linhas, e as categorias, nas colunas. Essa é outra maneira de apresentar uma matriz de formato indivíduos por variáveis, em que um indivíduo é caracterizado por uma e apenas uma modalidade de cada variável.

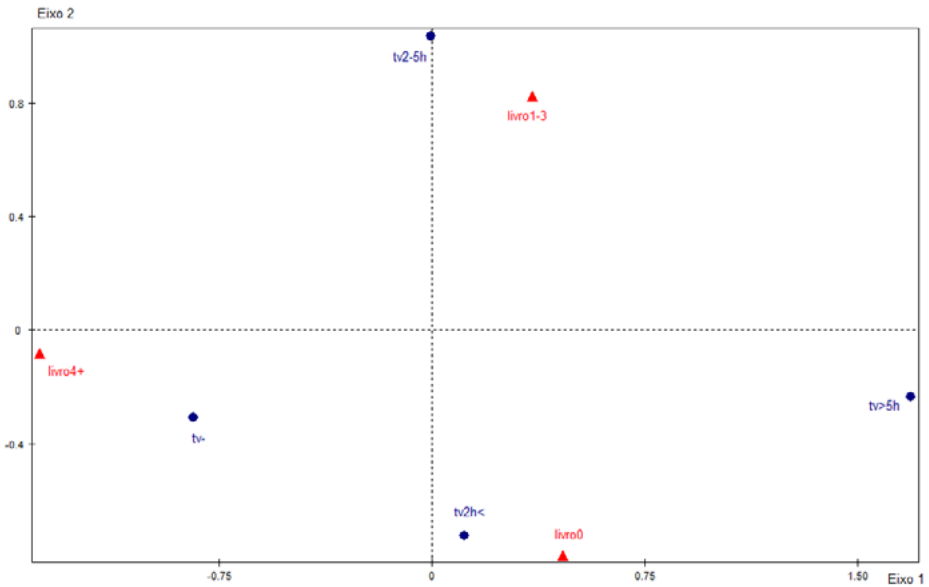
Tabela 1.2 | Matriz indicadora binária

	"livro0"	"livro1-3"	"livro4+"	"tv-"	"tv2h<"	"tv3-5h"	"tv>5h"	Total
1	1	0	0	0	0	0	1	2
2	0	0	1	1	0	0	0	2
3	0	1	0	0	1	0	0	2
N	1	0	0	1	0	0	0	2

Fonte: elaboração própria.

É importante notar que os resultados de uma ACM aplicada a duas variáveis são idênticos, em termos das localizações relativas das modalidades no plano fatorial, àquela de uma ACS aplicada a uma tabela de contingência construída a partir do cruzamento dessas mesmas variáveis, como é possível observar na Figura 1.2.

Figura 1.2 | Nuvem de modalidades resultante da ACM aplicada às variáveis "leitura de livros" e "audiência à TV", projetadas no plano cartesiano formado pelos eixos 1 e 2



Fonte: elaboração própria.

Principais conceitos: distância, inércia, contribuição

A **distância entre dois indivíduos é criada por questões para as quais eles "escolhem" categorias de resposta (modalidades) diferentes**. A distância entre dois indivíduos i e i' em uma determinada questão q é dada pela seguinte fórmula: $d_q^2(i,i') = \frac{1}{f_k} + \frac{1}{f_{k'}}$, em que k e k' são duas categorias de resposta à questão q; e f_k e $f_{k'}$ correspondem às proporções de indivíduos que "escolheram" tais categorias.[14] A distância total entre dois indivíduos é dada pela soma das distâncias ao quadrado em cada questão, dividida

[14] As notações matemáticas seguem o padrão utilizado em Le Roux e Rouanet (2004, 2010).

por $\frac{1}{Q}$, em que Q corresponde ao número total de questões (ou variáveis) ativas, como na fórmula a seguir: $d^2(i,i') = \frac{1}{Q}\sum_{q \in Q} d_q^2(i,i')$.

Em suma, **quanto mais dissimilares forem os perfis de respostas de dois indivíduos** (ou seja, escolherem categorias diferentes em um conjunto definido de questões), **maior será a distância entre eles – ou melhor, entre os pontos que os representam – na nuvem de indivíduos**.

Em relação à nuvem de modalidades, há dois aspectos que devem ser observados: a distância de uma categoria k em relação ao centro do espaço geométrico (ou **baricentro**) G; e as distâncias entre duas categorias quaisquer.[15] A primeira pode ser calculada através da seguinte fórmula: $d(k,G) = \sqrt{\frac{1}{f_k} - 1}$. Por definição, portanto, **quanto menor a frequência de uma categoria, mais distante estará do baricentro (G)**. Por sua vez, as distâncias entre duas categorias k e k' é dada por: $d(k,k') = \sqrt{\frac{1}{f_k} + \frac{1}{f_{k'}} - 2\frac{f_{kk'}}{f_k f_{k'}}}$, em que f_k é a proporção de indivíduos que escolheram a categoria k em dada variável; $f_{k'}$, a proporção dos que possuem o valor k'; e $f_{kk'}$, a proporção dos indivíduos que apresentam os dois valores (ou seja, escolheram as duas categorias). Por definição, então, se os valores f_k , $f_{k'}$ e $f_{kk'}$ forem idênticos, **a distância entre as categorias é nula**. Além disso, **quanto menor o valor de** $f_{kk'}$, ou seja, quanto menor a proporção de indivíduos que escolheram as duas categorias, **maior a distância entre elas**.

Chegamos, então, a algumas propriedades importantes da ACM:

> i. somente respostas (modalidades das variáveis) diferentes produzem distância entre os indivíduos e, **quanto menor a frequência das categorias de discordância, maior a distância entre os indivíduos**. Tal propriedade, **que aumenta a importância das categorias relativamente raras**, é desejável até certo ponto. Isso significa que categorias pouco frequentes (abaixo de 5%) devem ser recodificadas e agrupadas com outras ou, então, inseridas como **passivas** (voltaremos a este conceito adiante);
>
> ii. se um indivíduo escolhe categorias ou modalidades pouco frequentes (e, portanto, tem um perfil de respostas relativamente raro), **o ponto que o representa estará mais longe do centro da nuvem**;

[15] O baricentro equivale à média ponderada de pontos (Le Roux; Rouanet, 2010, p. 41).

iii. quanto mais duas categorias forem escolhidas pelos mesmos indivíduos, menor a distância relativa entre elas;

iv. quanto menos frequente for uma categoria, mais distante estará do centro da nuvem;

v. há uma perfeita correspondência entre a nuvem de indivíduos e a nuvem de categorias.

Na ACM, a **inércia** ou **variância** é um produto do número de variáveis e de suas categorias. É dada por: $\frac{K}{Q} - 1$, em que K é o número de categorias ativas; e Q, o número de questões ativas. Nesse caso, "a variância da nuvem não depende dos dados" (Le Roux; Rouanet, 2004, p. 187), mas das **escolhas referentes à codificação** (por exemplo: quantidade de categorias por questões) **e ao número de variáveis** que se deseja incluir. Quando todas as variáveis são **binárias, a inércia será sempre igual a 1**. Há duas implicações importantes aqui:

1. a inércia na ACM não representa uma medida da força da associação estatística em uma tabela, como se dá na ACS;

2. a **contribuição** das variáveis para a inércia total depende essencialmente da codificação, sendo dada pela seguinte fórmula: $Ctr_q = \frac{K_{q-1}}{K-Q}$, em que K_q é o número de categorias da questão q, K é o número total de categorias e Q, o número de questões. Isso significa que, ao construirmos um espaço (exemplo: um espaço dos gostos culturais), devemos buscar **um equilíbrio do número de categorias por questões e do número de questões por tópico investigado**.[16] Caso contrário, uma variável ou um conjunto de variáveis de um determinado tópico terá uma contribuição muito maior que a de outras(os) para a definição das distâncias nas nuvens.

Para entendermos a noção de **eixos** e os **valores próprios** (que correspondem à decomposição da inércia total), retomemos o exemplo do primeiro capítulo, em que temos 20 questões com duas categorias cada, totalizando, assim, 40 categorias. Para calcular o número total de padrões de respostas, basta multiplicar 2_{q1} x 2_{q2} x 2_{q3} x ... 2_{q20}, o que

[16] Cf. Hjellbrekke (2019, p. 35-36).

totaliza 1.048.576 possibilidades. Obviamente, não podemos representá-los em um plano tridimensional. Como vimos, a ACM (e a AGD mais amplamente) busca "redefinir as dimensões do espaço de modo que as principais dimensões capturem a maior variância possível, possibilitando descrições dos dados em menor dimensão" (Greenacre; Blasius, 2006, p. 5). Cada nuvem de pontos será referida por seus eixos principais. O número de eixos é dado pela seguinte fórmula: K - Q. "O primeiro eixo principal fornece o melhor ajuste unidimensional da nuvem... e a variância da nuvem projetada no eixo 1 é λ_1 [valor próprio ou autovalor]. De forma similar, o plano gerado pelos eixos principais 1 e 2... fornece o melhor ajuste bidimensional, com variância λ_1 e λ_2, e assim por diante" (Le Roux; Rouanet, 2010, p. 39).[17] Ocorre, portanto, uma **decomposição** da variância total ao longo dos eixos, cada um deles "explicando" uma porção dessa variância, em magnitudes decrescentes. **O objetivo é buscar o menor número de eixos ou dimensões que dê conta da maior parte da variância possível**. Ou seja, a questão, que discutiremos a seguir, é decidir quantos eixos reter para interpretação. Se, por exemplo, decidirmos reter três eixos, então bastaria construir duas nuvens de modalidades e duas de indivíduos, formadas pelos eixos 1 x 2 e 1 x 3, descrevendo as oposições ou contrastes ao longo dos três eixos mencionados.

Outra propriedade fundamental da ACM decorre da fórmula exposta anteriormente: **o número de eixos, ou seja, a dimensionalidade dos espaços depende inteiramente das estratégias de codificação dos dados**. "Como regra geral, quanto maior o número de categorias, menor será a porcentagem da inércia explicada por cada eixo." (Hjellbrekke, 2019, p. 37).

Por fim, consideremos a **contribuição** da modalidade, definida como a **proporção da variância devida à modalidade**. Como dito anteriormente, cada eixo "explica" uma proporção da variância ou inércia total. Por sua vez, a variância de cada eixo pode ser decomposta em termos das contribuições (proporcionais) de cada modalidade. A contribuição de uma modalidade k para a variância total é dada pela seguinte fórmula:

[17] Valor próprio é a tradução do termo *eigenvalue*.

$Ctr_k = \frac{1-f_k}{K-Q}$. Por sua vez, a contribuição da modalidade k para um eixo l pode ser calculada através da fórmula: $Ctr_k^l = \frac{\frac{f_k}{Q}(y_l^k)^2}{\lambda_l}$, em que f_k é a frequência relativa da modalidade; $\frac{f_k}{Q}$, o peso relativo; y_l^k, a coordenada da modalidade k no eixo l; e, por fim, λ_l, a variância do eixo l.[18] Por definição, portanto, a **contribuição da modalidade para o eixo** (que é, ao lado da coordenada da modalidade, a medida mais importante para a interpretação dos eixos) depende tanto de seu peso relativo quanto da coordenada, o que significa que, **quanto mais rara for a categoria e, portanto, mais distante estiver do centro do eixo, maior tenderá a ser a sua contribuição**. Depende, também, das escolhas de codificação: quanto maior o número de questões, menor será a contribuição de cada modalidade. Por fim, a contribuição de uma modalidade para a variância total é influenciada tanto pela frequência relativa quanto pela diferença entre a quantidade de modalidades e a quantidade de questões, o que significa que quanto maior o número de modalidades por questão menor tenderá a ser a contribuição de cada modalidade para a variância total.

Passos básicos de uma ACM

i. construir uma matriz de dados no formato indivíduos por variáveis;

ii. codificar adequadamente as variáveis, para que haja um equilíbrio razoável do número de modalidades por variável;

iii. escolher as variáveis **ativas** e **suplementares**, considerando um equilíbrio do número de variáveis e modalidades por tópico investigado;

iv. examinar se alguma modalidade apresenta frequência menor do que 5% e, então, agrupá-la a outra categoria ou, se não for possível, inseri-la como **passiva**;

[18] A contribuição de uma modalidade para um eixo é **relativa**, pois é obtida dividindo-se a porção da variância do eixo devida a um ponto pela variância do eixo (Le Roux; Roaunet, 2010, p. 29).

v. observar se há indivíduos com proporções elevadas de respostas ausentes em relação ao total de variáveis;

vi. decidir quantos eixos interpretar;

vii. inspecionar as duas nuvens e observar se há modalidades ou indivíduos relativamente muito distantes do centro das nuvens, tendendo a "esticá-las" excessivamente e prejudicando a visualização das oposições e afinidades entre os pontos. Caso isso ocorra, é aconselhável refazer a análise inserindo a modalidade ou indivíduo como passivo(a);

viii. interpretar os eixos retidos com base na nuvem de categorias;

ix. fazer a "análise dos dados estruturados", ou seja, apreender os determinantes das oposições descritas inserindo as variáveis suplementares;

x. examinar a nuvem de indivíduos, observando a dispersão das subnuvens resultantes da partição da nuvem global conforme as categorias de uma ou mais variáveis suplementares.

Comecemos com a **construção da tabela no formato indivíduos por variáveis**. A preparação do questionário deve considerar alguns princípios para a utilização da ACM, como o da **homogeneidade** (cada variável deve ser construída como uma questão com um número finito de categorias de resposta) e o da **exaustividade** (repertório significativo de variáveis para cada domínio sob investigação). Diferentemente de outras técnicas multivariadas, a ACM (mais amplamente, a AGD) não impõe limites muito estritos no que se refere à parcimônia: ao contrário, essa técnica possibilita mesmo revelar padrões complexos em bases de dados com muitas variáveis. Embora não exista consenso na literatura quanto à necessidade de um "n" mínimo, aceita-se geralmente a recomendação de, pelo menos, uma equivalência entre o número de variáveis e o tamanho da amostra (Di Franco, 2016).[19] Além disso, a operacionalização dos conceitos deve considerar que a ACM funciona melhor quando produzimos **diferenciações qualitativas** do fenômeno investigado, ou seja, quando temos um bom número de variáveis nominais, cujas categorias não podem ser ordenadas (pelo menos, não de forma direta).

[19] Obviamente, o uso de fontes secundárias nos impõe dificuldades adicionais.

Após a realização de análises estatísticas elementares (frequências, cruzamentos bivariados, medidas de dispersão etc.), passa-se, então, à **definição das variáveis e indivíduos ativos e suplementares**. Variáveis (e modalidades) **ativas** são aquelas consideradas para a construção dos eixos, ou seja, para a definição das distâncias relativas na nuvem de modalidades. Na ACM, as variáveis ativas devem ser necessariamente **categóricas** ou **categorizadas** (na ACP, as variáveis ativas são **quantitativas**). O entendimento é similar para os indivíduos **ativos**, quer dizer, são aqueles cujas respostas às questões (as escolhas das modalidades de resposta) são consideradas para a construção das distâncias relativas na nuvem de indivíduos. Variáveis **suplementares**, por sua vez, não participam da determinação dos eixos ou da definição das distâncias. Podem, no entanto, ser **projetadas** nas nuvens de modalidades para o propósito da "análise de dados estruturados", que possibilita apreender os **fatores estruturantes** das oposições ou afinidades reveladas nos diferentes espaços relacionais (voltarei a esse ponto adiante). Diferentemente das variáveis ativas, as suplementares podem ser **quantitativas**.

Como definir quais variáveis serão inseridas como ativas e quais serão suplementares? Deixando de lado, por ora, a questão dos indivíduos, podemos afirmar que a resposta a essa questão depende essencialmente dos **interesses de pesquisa**. Como regra geral, **variáveis de diferentes tipos não devem ter o mesmo estatuto**. Assim, num questionário sobre consumo cultural, que possua itens sobre participação, gosto e conhecimento sobre diversos gêneros culturais, e também um conjunto de indicadores sobre as propriedades sociais dos indivíduos (exemplos: idade, sexo, raça/cor, ocupação, renda, escolaridade, religião, origem etc.), se as variáveis referentes a aspectos das práticas culturais forem consideradas, conforme os propósitos da pesquisa, como ativas, então as que se referem ao perfil sociodemográfico serão inseridas como suplementares. No exemplo mencionado, partiríamos, digamos, da construção de "mapas culturais" para, então, buscar "explicar" os contrastes observados pela projeção das modalidades indicadoras das propriedades sociais. Ao invés, poderíamos partir do **espaço social**, noção emprestada da sociologia bourdieusiana para dar conta da estrutura de distribuição de recursos ou capitais entre os agentes, construído a

partir da inserção de tais modalidades como **ativas** e, então, projetar as variáveis referentes às práticas como suplementares (a nuvem de modalidades assim construída funcionaria como uma espécie de "mapa preditivo" das práticas).

Ainda no que se refere a escolhas preliminares, é essencial que exista um **equilíbrio do número de categorias das variáveis ativas** e também do **número de variáveis por diferentes temas ou domínios** que se deseja investigar. Vimos que o número de categorias afeta diretamente a **contribuição** de cada variável para a construção das nuvens. Por certo, é possível alcançar um equilíbrio relativo por meio da recodificação de variáveis, mas é importante que tal requisito seja considerado já na etapa referente à construção do questionário.[20]

A **questão sobre quantos eixos** reter para interpretação segue a seguinte lógica: busca-se "interpretar o menor número possível de eixos, mas tantos quantos forem necessários para reter todas as informações relevantes na tabela" (Hjellbrekke, 2019, p. 18).

Os critérios que orientam a escolha dos eixos para interpretação são os seguintes:

i. **A redução da diferença entre os autovalores (ou valores próprios).** A diferença entre λ_1 e λ_2, dada pela fórmula $\frac{\lambda_1 - \lambda_2}{\lambda_1}$, é igual a 32,9%; entre λ_2 e λ_3, 46,6%; λ_3 e λ_4, 23,2% e, então, cai acentuadamente: entre λ_4 e λ_5, 8,3%; entre λ_5 e λ_6, 7,5% (ver Tabela 1.3). Por tal critério, reteríamos 4 eixos.[21]

ii. **A proporção** modificada **da variância acumulada**. Em função da elevada dimensionalidade das nuvens, as taxas de variância dos primeiros eixos tendem a ser fortemente **subestimadas**, sendo geralmente muito baixas. Por isso, Jean-Paul Benzécri propôs uma fórmula para corrigir tal subestimação[22]. Considerando as taxas modificadas, vemos que o primeiro eixo "explica" quase

[20] Para um exemplo desse ponto, ver Bennett *et al*. (2009, p. 45).

[21] Os valores mencionados referem-se a um exercício diferente, com mais variáveis e modalidades, que será tratado a seguir.

[22] Cf. Le Roux e Rouanet (2010, p. 39).

70% da variância total; e os dois primeiros, quase 95% dela. Se acrescentarmos o terceiro, chegamos a pouco mais de 98% da variância total. Lembremos que tais proporções variarão conforme o número de variáveis e categorias ativas: quanto maior o número delas e, sobretudo, quanto mais categorias em relação a questões ativas, maior será o número de eixos e, portanto, menor tenderá a ser a proporção da variância "explicada" pelos eixos iniciais. Com base nesse critério, poderíamos reter 2 eixos.

iii. **O teste de *Cattell* [*scree test*].** O teste ordena os eixos conforme a porcentagem da inércia explicada. Retêm-se aqueles à esquerda do ponto em que a curva começa a achatar-se. Observando a Figura 1.3, percebemos que é a partir do 4º eixo (depois, mais fortemente, a partir do 7º) que ocorre o achatamento da curva. Por isso, considerando tal critério, reteríamos não mais do que 4 eixos.

iv. **Por fim, mas não menos importante, a interpretabilidade dos eixos** (Le Roux; Rouanet, 2010, p. 102). Embora, pelo segundo critério, não aumente muito significativamente a taxa de variância explicada, a inclusão do terceiro eixo possibilitou apreender um contraste nos padrões de gosto observável em outros estudos já realizados sobre o tema, como veremos. Por isso, considerando todos os critérios enumerados anteriormente, optou-se por reter três eixos para interpretação. O quarto eixo, de fato, acrescenta muito pouco à variância total, conforme a taxa modificada (menos de 1%).

Tabela 1.3 | Autovalores, porcentagem da variância e porcentagem acumulada dos cinco primeiros eixos da ACM

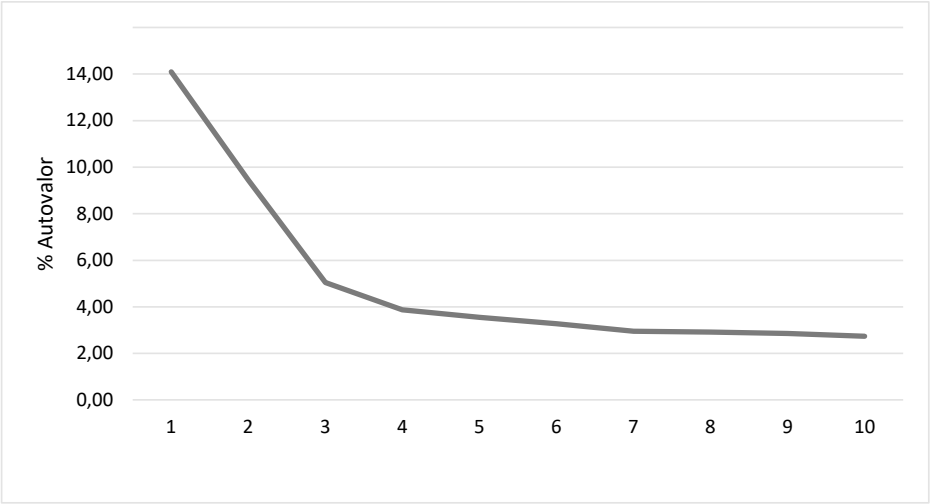
Eixo	Autovalor	% da variância (inércia)	% da variância (acumulada)	Autovalor modificado	% variância modificada
1	0,1665	14,09	14,09	0,0197	69,98
2	0,1116	9,44	23,53	0,007	24,91
3	0,0596	5,04	28,57	0,0009	3,23
4	0,0457	3,87	32,44	0,0003	0,89
5	0,0419	3,55	35,99	0,0001	0,51

Fonte: elaboração própria.

Uma vez decidido o número de eixos a reter, passemos, então, à questão sobre como interpretá-los. Conforme nos ensina Benzécri (1992, p. 405; *apud* Le Roux; Rouanet, 2004),

> interpretar um eixo implica descobrir o que é similar, de um lado, entre todos os elementos que estão à direita [ou acima] da origem e, de outro, entre todos os elementos que estão à esquerda [ou abaixo] dela; e expressar, com concisão e precisão, o contraste (ou oposição) entre os dois extremos (p. 49).

Figura 1.3 | Teste de Cattell



Fonte: elaboração própria.

Interpretar um eixo, portanto, consiste em apreender as oposições ou contrastes entre conjuntos de pontos ou modalidades. Para tanto, consideramos as **coordenadas** das categorias cujas **contribuições** – lembremos que a contribuição de um ponto para um eixo é uma medida estatística que depende **tanto da distância do ponto em relação à origem quanto de seu peso** – para a variância do eixo excedem a média, conforme $\frac{1}{K}$, e também as questões/variáveis cujas contribuições acumuladas para o eixo são maiores do que $\frac{1}{Q}$. É possível também considerar o critério da **qualidade da representação**, que é uma medida

de quanto um eixo contribui para um ponto ou modalidade (Le Roux; Rouanet, 2010, p. 27). No entanto, devido à elevada dimensionalidade da ACM, a qualidade de representação expressará "em muitas situações, uma visão muito conservadora da capacidade descritiva de um dado plano fatorial" (Hjellbrekke, 2019, p. 39). Por isso, não é comum considerar a qualidade da representação como uma informação relevante para a interpretação dos resultados da ACM.

Um exemplo empírico: consumo cultural na região metropolitana de Belo Horizonte

No exemplo a seguir, são investigados padrões de consumo cultural considerando como ativas as variáveis que mensuram a **participação** de indivíduos em eventos culturais fora de casa, **gostos** e **rejeições por gêneros televisivos e musicais**, além da **frequência com que leem jornal**, **assistem à televisão** e **ouvem rádio**. São 33 variáveis ativas – que, como sabemos, devem ser categóricas ou, pelo menos, categorizadas – e 72 modalidades ativas. Buscou-se construir um equilíbrio no número de categorias por variáveis (que varia entre 2 e 4) e também de variáveis por tópico (40 modalidades de gosto e 32 modalidades de participação ou consumo, de tipo ativo ou receptivo). Temos um total de 39 eixos. Desses, reteremos apenas 3 para interpretação – os critérios para a escolha estão explicitados acima. **A interpretação é feita, de preferência, para cada eixo separadamente, e não por quadrantes**.[23] Para tanto, serão construídas três nuvens de modalidades, cada uma exibindo as modalidades que mais contribuem para cada um dos três eixos retidos.

[23] No último capítulo, apresentarei uma exceção a essa regra que orienta a interpretação dos eixos.

Tabela 1.4 | Coordenadas e contribuições das modalidades ativas para os três primeiros eixos da ACM

Variáveis / modalidades	Coordenadas			Contribuições		
	Eixo 1	Eixo 2	Eixo 3	Eixo 1	Eixo 2	Eixo 3
Quais músicas você gosta ou não gosta de ouvir?						
CC1A						
romântica+	-0,01	0,37	-0,18	0,00	2,51	1,12
romântica-	0,03	-0,79	0,39	0,01	5,38	2,41
TOTAL				0,01	7,89	3,53
CC1B						
rock+	-0,69	0,23	0,53	2,90	0,51	4,80
rock-	0,35	-0,12	-0,27	1,49	0,26	2,47
TOTAL				4,39	0,77	7,28
CC1C						
clássica+	-0,43	0,27	-0,56	1,34	0,77	6,20
clássica-	0,28	-0,17	0,36	0,86	0,50	3,98
TOTAL				2,21	1,27	10,18
CC1D						
sertaneja+	0,18	0,30	-0,10	0,40	1,62	0,33
sertaneja-	-0,38	-0,63	0,21	0,86	3,44	0,69
TOTAL				1,26	5,07	1,02
CC1E						
rap+	-0,17	0,81	1,38	0,08	2,62	14,02
rap-	0,03	-0,14	-0,23	0,01	0,45	2,39
TOTAL				0,09	3,07	16,41
CC1F						
funk+	-0,11	0,90	1,52	0,03	2,92	15,64

Variáveis / modalidades	Coordenadas			Contribuições		
	Eixo 1	Eixo 2	Eixo 3	Eixo 1	Eixo 2	Eixo 3
Quais músicas você gosta ou não gosta de ouvir?						
funk-	0,02	-0,14	-0,23	0,00	0,45	2,40
TOTAL				0,03	3,37	18,04
CC1G						
mpb+	-0,48	0,36	0,00	2,25	1,85	0,00
mpb-	0,56	-0,41	0,00	2,61	2,14	0,00
TOTAL				4,86	3,99	0,00
CC1H						
samba+	-0,14	0,66	-0,02	0,15	5,50	0,01
samba-	0,11	-0,56	0,02	0,13	4,67	0,01
TOTAL				0,28	10,17	0,03
CC1I						
pagode+	0,14	0,67	0,23	0,14	5,11	1,18
pagode-	-0,10	-0,49	-0,17	0,11	3,72	0,86
TOTAL				0,25	8,83	2,05
CC1J						
religiosa+	0,26	0,24	-0,38	0,74	0,88	4,19
religiosa-	-0,37	-0,33	0,53	1,04	1,24	5,91
TOTAL				1,78	2,12	10,09
CC1K						
jazz+	-0,78	0,64	-0,22	2,44	2,48	0,55
jazz-	0,22	-0,18	0,06	0,69	0,71	0,16
TOTAL				3,13	3,19	0,71

Variáveis / modalidades	Coordenadas			Contribuições		
	Eixo 1	Eixo 2	Eixo 3	Eixo 1	Eixo 2	Eixo 3
Programas de televisão que gosta e não gosta de assistir:						
CC3A						
novela+	0,15	0,28	0,01	0,27	1,32	0,01
novela-	-0,24	-0,44	-0,02	0,41	2,05	0,01
TOTAL				0,68	3,37	0,02
CC3B						
variedades+	0,14	0,36	-0,07	0,24	2,28	0,15
variedades-	-0,26	-0,65	0,12	0,44	4,08	0,27
TOTAL				0,68	6,36	0,43
CC3C						
esporte+	-0,12	0,37	0,19	0,14	1,83	0,86
esporte-	0,12	-0,35	-0,18	0,13	1,74	0,82
TOTAL				0,27	3,57	1,69
CC3D						
filmes+	-0,14	0,30	0,09	0,24	1,57	0,29
filmes-	0,28	-0,58	-0,18	0,48	3,08	0,57
TOTAL				0,72	4,65	0,85
CC3E						
jornais+	-0,02	0,14	-0,10	0,01	0,45	0,43
jornais-	0,14	-0,99	0,71	0,04	3,26	3,11
TOTAL				0,05	3,71	3,54
CC3F						
doc+	-0,34	0,22	-0,18	1,37	0,87	1,15
doc-	0,68	-0,44	0,37	2,77	1,76	2,32

Variáveis / modalidades	Coordenadas			Contribuições		
	Eixo 1	Eixo 2	Eixo 3	Eixo 1	Eixo 2	Eixo 3
Programas de televisão que gosta e não gosta de assistir:						
TOTAL				4,14	2,63	3,47
CC3G						
religiosos+	0,38	0,27	-0,49	1,16	0,84	5,45
religiosos-	-0,30	-0,21	0,39	0,92	0,67	4,31
TOTAL				2,08	1,51	9,76
CC3H						
humor+	-0,08	0,44	0,04	0,07	3,14	0,04
humor-	0,11	-0,64	-0,06	0,09	4,55	0,06
TOTAL				0,16	7,70	0,11
CC3I						
musical+	-0,15	0,46	-0,07	0,26	3,48	0,16
musical-	0,23	-0,68	0,11	0,38	5,14	0,23
TOTAL				0,64	8,61	0,39
Foi a um show musical nos últimos dois anos?						
CC5A						
show+ (sim)	-0,68	-0,01	0,18	3,98	0,00	0,77
show-(não)	0,60	0,00	-0,16	3,52	0,00	0,68
TOTAL				7,50	0,00	1,46
Foi a um concerto de orquestra nos últimos dois anos?						
CC5B						
concerto+	-1,28	-0,07	-0,31	3,42	0,02	0,57
concerto-	0,16	0,01	0,04	0,44	0,00	0,07
TOTAL				3,85	0,02	0,64

Variáveis / modalidades	Coordenadas			Contribuições		
	Eixo 1	Eixo 2	Eixo 3	Eixo 1	Eixo 2	Eixo 3
Foi a uma exposição de arte nos últimos dois anos?						
CC5C						
exposição+	-1,07	-0,27	-0,14	6,54	0,64	0,31
exposição-	0,49	0,13	0,06	3,03	0,29	0,15
TOTAL				9,57	0,93	0,46
Foi a um espetáculo de teatro ou de dança nos últimos dois anos?						
CC5D						
espetáculo+	-1,00	-0,03	-0,11	5,89	0,01	0,18
espetáculo-	0,47	0,02	0,05	2,78	0,00	0,09
TOTAL				8,67	0,01	0,27
Foi a um centro cultural nos últimos dois anos?						
CC5E						
ccultural+	-1,07	-0,10	-0,08	5,35	0,07	0,09
ccultural-	0,37	0,03	0,03	1,85	0,02	0,03
TOTAL				7,20	0,10	0,12
Foi a um monumento de valor histórico-cultural nos últimos dois anos?						
CC5F						
monumento+	-0,68	-0,09	-0,08	4,06	0,11	0,16
monumento-	0,62	0,08	0,07	3,70	0,10	0,15
TOTAL				7,76	0,22	0,31
Foi a um lugar com música ao vivo nos últimos dois meses?						
CC6B						
maovivo+	-0,60	-0,02	0,17	3,15	0,01	0,74
maovivo-	0,56	0,02	-0,16	2,95	0,01	0,69
TOTAL				6,10	0,01	1,43

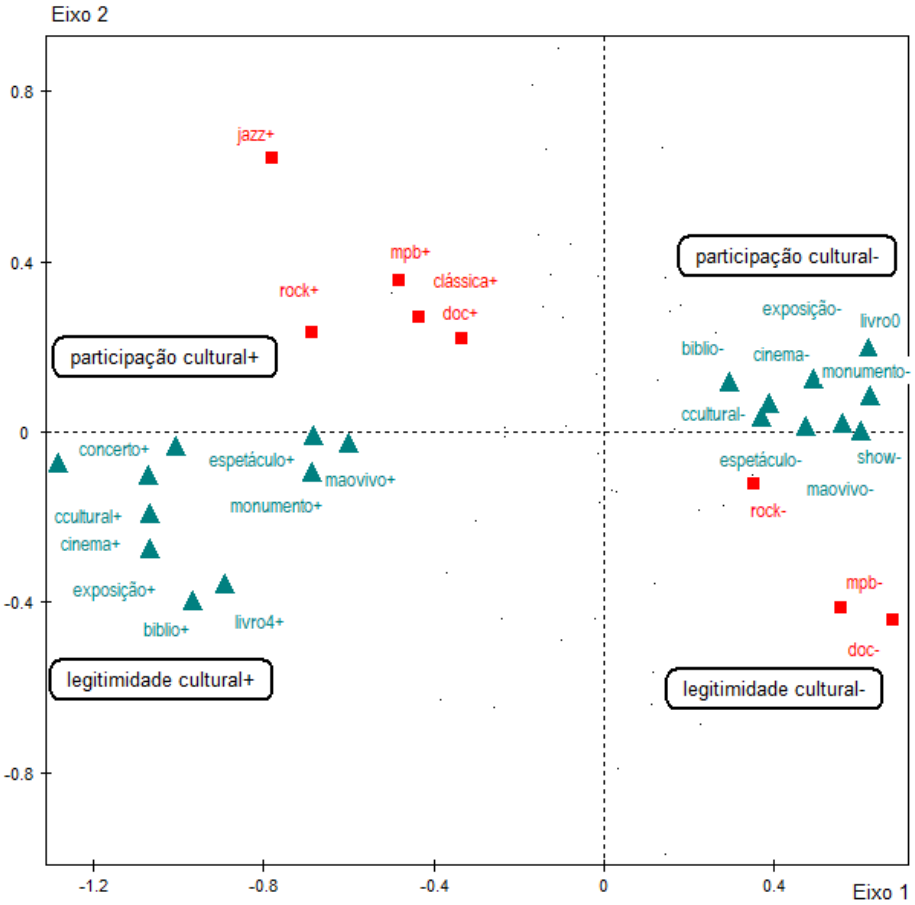
Variáveis / modalidades	Coordenadas			Contribuições		
	Eixo 1	Eixo 2	Eixo 3	Eixo 1	Eixo 2	Eixo 3
Foi ao cinema nos últimos dois meses?						
CC6C						
cinema+	-1,07	-0,19	0,00	5,53	0,27	0,00
cinema-	0,39	0,07	0,00	2,02	0,10	0,00
TOTAL				7,54	0,36	0,00
Foi a uma biblioteca nos últimos dois meses?						
CC6D						
biblio+	-0,97	-0,39	-0,17	3,97	0,98	0,33
biblio-	0,29	0,12	0,05	1,21	0,30	0,10
TOTAL				5,18	1,28	0,43
Com que frequência você assiste à televisão?						
Tv						
tv>5h (mais de 5 horas por dia)	0,31	0,45	0,24	0,22	0,69	0,37
tv2-5h (entre duas e cinco horas por dia)	-0,10	0,27	-0,05	0,05	0,62	0,03
tv2h< (até duas horas por dia)	-0,01	-0,05	0,05	0,00	0,02	0,04
tv- (algumas vezes por semana ou nunca)	-0,02	-0,44	-0,11	0,00	1,50	0,17
TOTAL				0,28	2,82	0,62
Quantos livros leu nos últimos 12 meses?						
Livros						
livro0 (nenhum)	0,62	0,20	0,23	2,69	0,40	1,05

Variáveis / modalidades	Coordenadas			Contribuições		
	Eixo 1	Eixo 2	Eixo 3	Eixo 1	Eixo 2	Eixo 3
livro1-3 (um a três)	-0,09	0,01	-0,08	0,06	0,00	0,12
livro4+ (quatro ou mais)	-0,89	-0,36	-0,26	3,28	0,78	0,78
TOTAL				6,03	1,18	1,95
Com que frequência você lê jornal?						
Jornal						
jornal++ (todos os dias ou quase)	-0,50	0,33	-0,13	0,73	0,46	0,13
jornal+ (algumas vezes na semana)	-0,24	-0,01	-0,06	0,37	0,00	0,06
jornal- (quase nunca ou nunca)	0,35	-0,10	0,09	1,06	0,13	0,18
TOTAL				2,15	0,60	0,37
Com que frequência você ouve rádio?						
Rádio						
rádio>5h (mais de 5 horas por dia)	0,21	0,22	0,20	0,15	0,26	0,41
rádio2-5h (entre duas e cinco horas por dia)	0,13	0,12	0,14	0,05	0,07	0,17
rádio2h< (até duas horas por dia)	-0,23	0,01	0,18	0,24	0,00	0,38
rádio- (algumas vezes por semana ou nunca)	-0,01	-0,16	-0,26	0,00	0,29	1,41
TOTAL				0,44	0,62	2,37

Fonte: elaboração própria.

Na Figura 1.4, temos o plano construído pelos eixos 1 e 2, **sendo exibidas as modalidades que mais contribuem para o primeiro eixo**. Para interpretação dos eixos, reporta-se inicialmente o número de questões e modalidades com contribuições acima da média, conforme as fórmulas exibidas anteriormente: a média das contribuições das questões é 3,03% ($\frac{1}{Q}$); das modalidades, 1,39% ($\frac{1}{K}$). A descrição do eixo é feita considerando, como mencionei antes, os contrastes. Por isso, é importante **nomeá-los**, de modo a tornar a interpretação mais inteligível. As contribuições – lembremos que os valores das contribuições são proporções da variância de cada eixo – das diferentes modalidades podem ser vistas na Tabela 1.4. As contribuições das variáveis podem ser calculadas simplesmente somando as contribuições de suas modalidades.

Figura 1.4 | Categorias que mais contribuem para o eixo 1, projetadas no plano fatorial formado pelos eixos 1 e 2



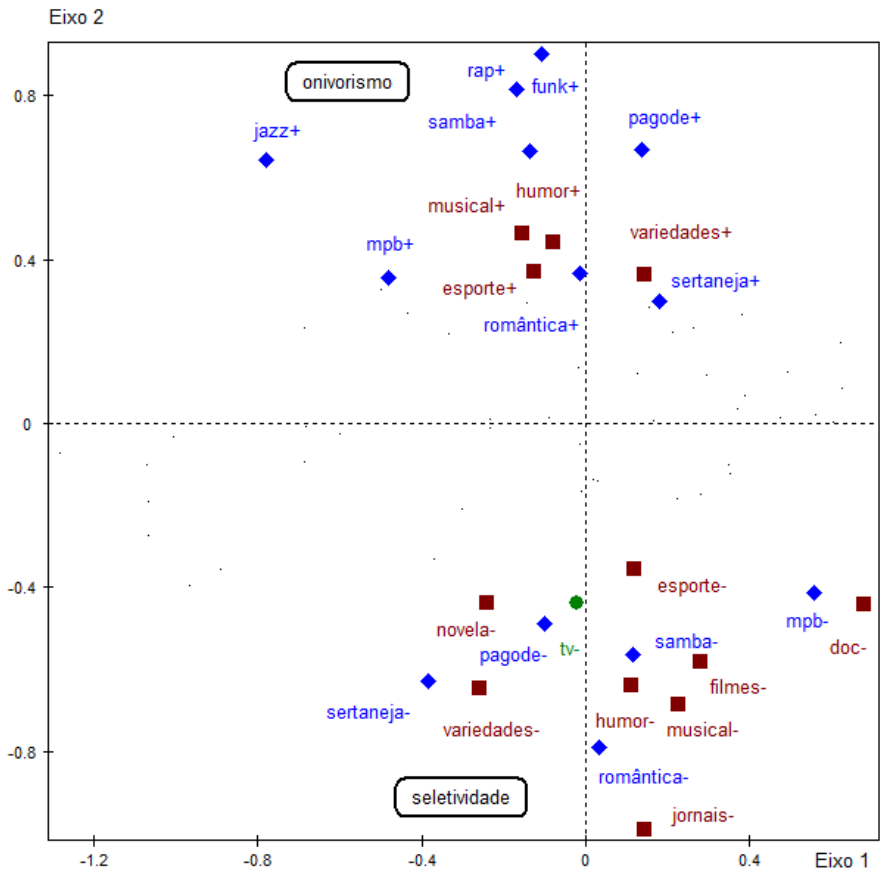
Fonte: elaboração própria.

A leitura da Tabela 1.4 indica que não há variáveis ou modalidades com contribuições singulares muito superiores às demais. A modalidade com maior contribuição proporcional é “exposição+”, que “explica” pouco mais de 6,5% da variância, que somada à “exposição-”, alcança 9,57%, que é a maior contribuição de uma variável para o eixo. Tais valores não são tão discrepantes em relação aos de outras variáveis e modalidades com contribuições acima da média. A inspeção da nuvem de modalidades

não revela nenhuma modalidade cujo ponto no plano cartesiano esteja relativamente muito distante do centro e daqueles de outras modalidades. Ou seja, a referida nuvem não aparenta estar "esticada" demais por uma ou outra modalidade, o que significaria que as demais modalidades estariam concentradas em uma região próxima ao centro.

A partir da Tabela 1.4, vemos que 14 variáveis e 27 modalidades possuem contribuições acima da média para o eixo horizontal (1), sendo 10 variáveis e 19 modalidades de consumo cultural que, com exceção de leitura de livros, referem-se à participação cultural fora de casa (por exemplo ir a um show, concerto, exposição, bar de música ao vivo etc.). Essas modalidades referentes a "saídas culturais" contribuem com aproximadamente 63% da variância do eixo. Temos, portanto, um eixo que opõe, sobretudo, engajamento *versus* desengajamento cultural (em termos de níveis de participação ou consumo cultural fora de casa). À esquerda do eixo horizontal (1), temos as modalidades que indicam a ida a shows, concertos, exposições, espetáculos, centros culturais, monumentos, bares com música ao vivo, cinema, biblioteca e também a leitura de quatro livros ou mais nos últimos seis meses. Nessa parte do eixo, ainda estão localizadas as modalidades referentes ao gosto por música clássica, MPB, *jazz* e *rock*, além de documentários, entre os gêneros televisivos, o que significa que o maior engajamento cultural está, de algum modo, relacionado com gostos culturais específicos. Do outro lado do eixo, estão as modalidades que indicam a não participação cultural, a ausência de leitura de livros e também a rejeição ao *rock*, à MPB e a documentários.

Figura 1.5 | Categorias que mais contribuem para o eixo 2, projetadas no plano fatorial formado pelos eixos 1 e 2



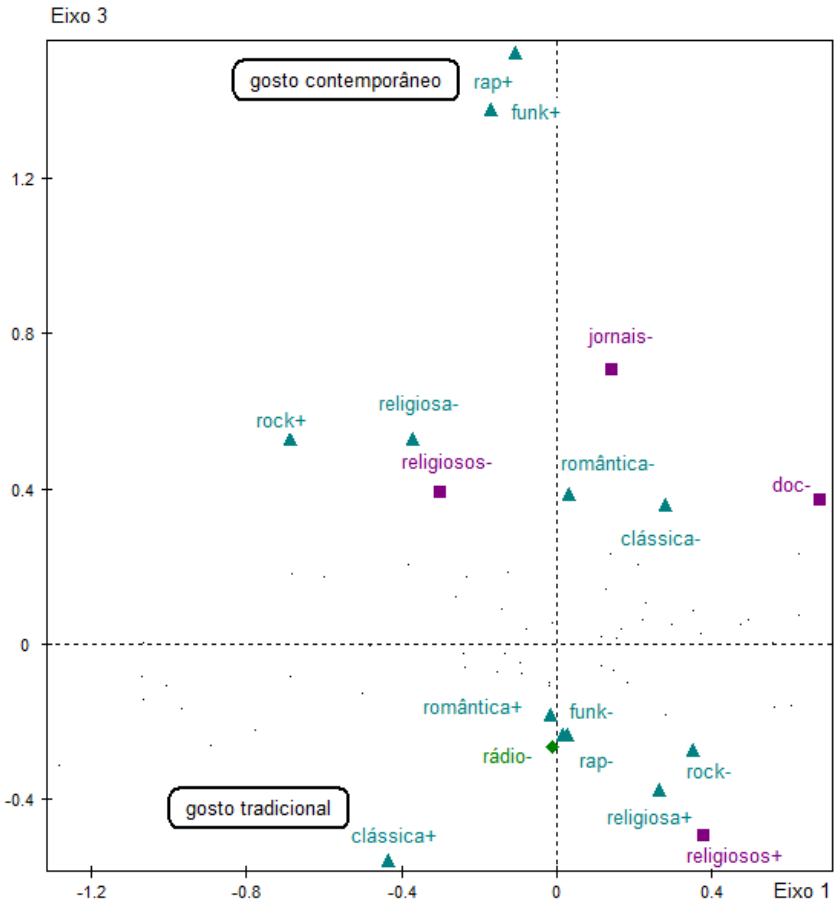
Fonte: elaboração própria.

Diferentemente, o eixo vertical (2), que pode ser visualizado na Figura 1.5, opõe predominantemente preferências por gêneros musicais e televisivos. Das 15 variáveis e 28 modalidades com contribuições acima da média para o eixo, 8 variáveis e 13 modalidades se referem a gostos musicais; 7 variáveis e 14 modalidades, a gostos televisivos; por fim, a modalidade que indica a mais baixa frequência de audiência à TV ("tv0"): 45,5% da variância do eixo é "explicada" pelas variáveis de

gosto musical e quase 38%, pelas de gosto televisivo. Na parte superior do eixo, estão as modalidades que indicam gosto por diversos gêneros musicais: romântica, sertaneja, *rap*, *funk*, MPB, pagode, *jazz*; e televisivos: novelas, programas de variedades, esportes, humor, filmes e musicais. Tal combinação de preferências expressa o que uma parte da literatura denominaria por gosto "onívoro", que "atravessa" certas fronteiras culturais, como aquela entre *jazz* e sertanejo, ou entre MPB e *funk*.[24] Abaixo, estão as modalidades que indicam rejeição a diversos gêneros musicais (romântica, sertaneja, MPB, samba, pagode) e televisivos (novela, variedades, esportes, filmes, jornais, documentários, humor, musicais) e também a modalidade que indica pouca ou inexistente audiência à TV. Trata-se aqui, portanto, de uma oposição entre gostos "onívoros", de um lado, e gostos "seletivos", de outro.

[24] Peterson (2005). Por não ser o objetivo do livro, não me aprofundarei em algumas implicações teóricas dos resultados da ACM, sobretudo no que se refere ao tema da estratificação social das práticas culturais. Para um aprofundamento nessa questão, Bertoncelo (2019b).

Figura 1.6 | Categorias que mais contribuem para o eixo 3, projetadas no plano fatorial formado pelos eixos 1 e 3



Fonte: elaboração própria.

Por fim, o terceiro eixo (vertical), que "explica" pouco mais que 3% da variância total, exibe um contraste entre o "gosto tradicional" e o "gosto contemporâneo" (Figura 1.6). Das 9 variáveis e 17 modalidades com contribuições acima da média para a inércia do eixo, 6 variáveis e 12 modalidades referem-se a gêneros musicais, que "explicam" mais de 65%

da inércia. De um lado, observamos um padrão de gosto que poderíamos denominar de "clássico" e que inclui preferências por música romântica, clássica e religiosa, e por programas religiosos na TV, além de rejeição ao *rock*, *rap* e *funk*. Do outro, um padrão de gosto contemporâneo, composto por preferências pelo *rock*, *rap* e *funk* e pela rejeição à música romântica, clássica, religiosa; por jornais, documentários e programas religiosos, entre os gêneros televisivos.

Outra informação importante para caracterizar as principais oposições nos eixos é aquela sobre **a contribuição interna dos desvios entre duas categorias ativas de uma mesma variável**: "para cada questão, nós também avaliamos a contribuição do desvio entre o baricentro das categorias selecionadas localizadas de um lado do eixo e o baricentro daquelas localizadas do outro lado, e expressamos essa contribuição como uma proporção da contribuição da questão." (Le Roux; Roaunet, 2010, p. 57)

Para calcular a contribuição interna, seguimos os seguintes passos (Hjellbrekke, 2019, p. 50):

i. **Identificamos as variáveis/questões mais importantes para um eixo e, então, as categorias com maiores contribuições relativas**. Em nosso exemplo, a variável "livros" tem duas de suas três categorias com contribuições acima da média. A partir da contribuição relativa da variável para a inércia do eixo (0,0603 ou 6,03%), calcula-se a **contribuição absoluta** com base na seguinte fórmula: $Cta_{kl} = \lambda \, x \, Ctr_q^l$, em que Ctr_q^l é a contribuição relativa da questão q para o autovalor (λ) do eixo l.[25] Temos, então, que a contribuição absoluta é: 0,1665 x 0,0603, que é igual a 0,01004.

ii. **Para se obter a contribuição interna resultante da oposição entre duas ou mais categorias dessa questão/variável**, calculamos:

[25] A contribuição relativa da questão para a variância de um eixo é obtida pela soma das contribuições relativas das categorias dessa questão para o referido eixo.

a. a **frequência absoluta do dipolo ($\tilde{n}_d$)**, dada por $\frac{1}{(\frac{1}{n_k}+\frac{1}{n_{k\prime}})}$, em que n_k é a frequência absoluta da modalidade *k* e $n_{k'}$, aquela da modalidade *k'*;

b. e, então, **o desvio (d) entre os baricentros** das duas modalidades, com base na seguinte fórmula: $d = y_l^k - y_l^{k\prime}$, em que y_l^k é a coordenada de *k* no eixo *l*, e, por fim, $y_l^{k\prime}$, a coordenada de *k'* no eixo *l*;

c. daí, **a contribuição absoluta do desvio**, utilizando a fórmula a seguir: $\frac{\tilde{n}_d}{n \times Q} x\ (d)^2$. No exemplo em questão, o valor resultante da aplicação dessa fórmula é: 0,0098.

iv. **Dividimos, por fim, esse valor por aquele calculado no primeiro passo**, obtendo, então, 0,976. Ou seja, 97,6% da contribuição da variável "livro" para o primeiro eixo decorre do desvio entre as duas categorias "livro0" e "livro4+".[26]

Análise de correspondências múltiplas específica

Essa variante da ACM permite desconsiderar modalidades de variáveis ativas na determinação dos eixos e no cálculo das distâncias nas nuvens de indivíduos e de categorias. Tais modalidades, denominadas de **passivas**, não são propriamente excluídas da análise, pois é possível projetá-las nos planos fatoriais. **Pontos ou categorias passivas ou suplementares assemelham-se por serem pontos sem massa ou peso**. Por definição, portanto, não contribuem para a inércia total ou para a inércia dos eixos. Conforme argumenta Hjellbrekke (2019, p. 57), "se o ponto não tem contribuição para nenhum eixo, não exercerá qualquer 'força' na direção de quaisquer eixos através do espaço multidimensional". Tomando o exemplo muito ilustrativo dado pelo autor, podemos comparar as

[26] Tal exercício é mais produtivo quando trabalhamos com variáveis com um número maior de categorias com contribuições significativas para a variância de um eixo.

categorias a bolas magnéticas e os eixos a hastes metálicas que se equilibram no centro de um espaço multidimensional. As categorias ativas funcionam como tais bolas magnetizadas, com forças desiguais para puxar as hastes em diferentes direções. As categorias passivas ou suplementares estão "desmagnetizadas", portanto, incapazes de exercer qualquer efeito sobre a direção das hastes (ou eixos).

Conceitualmente, é comum diferenciar as modalidades **passivas**, por um lado, como categorias de variáveis ativas, das modalidades **suplementares**, por outro, que são as categorias de variáveis projetadas após a construção das nuvens para permitir apreender os fatores estruturantes das oposições ou contrastes em cada eixo.

Quais são os critérios para definir uma categoria como **passiva**? Podemos indicar algumas situações típicas:

i. **Quando uma categoria tiver frequência menor do que 5%**. Lembremos que, quanto mais rara for uma categoria, mais distante estará do centro da nuvem de modalidades e, quanto mais raras forem as categorias de discordância, maior será a distância entre os indivíduos. Essa propriedade da ACM, **que tende a aumentar a importância das categorias pouco frequentes**, é desejável até certo ponto. Deve-se, inicialmente, buscar agrupar tais categorias pouco frequentes com outras categorias da mesma variável e, somente quando isso não for possível, inseri-las como passivas.

ii. **Quando uma categoria exibir uma contribuição relativamente muito elevada para um eixo retido para interpretação**. Ao reinseri-la como passiva, as demais categorias passarão a exibir contribuições maiores, possibilitando examinar mais adequadamente a estrutura de afinidades e contrastes entre elas.

iii. **Quando é necessário testar a estabilidade de uma solução**. É possível inserir diferentes categorias como passivas a cada vez e observar se os resultados diferem de modo significativo.

iv. **Quando um ponto está relativamente muito distante do centro da nuvem**. Pontos localizados nas extremidades de um eixo tendem a "esticar" demais as nuvens, dificultando a visualização dos contrastes.

v. **Quando há dados ausentes ou categorias sem pertinência analítica do ponto de vista da questão de pesquisa** (ex.: "outros"; "não sabe" etc.).[27]

Os passos para a ACM Específica são os mesmos que aqueles já descritos para a ACM, com duas diferenças:

i. para a construção das nuvens, é preciso definir quais são as categorias ativas, passivas e suplementares;

ii. é importante descartar os indivíduos cujas respostas se enquadram nas categorias definidas como passivas em mais de 1/5 das questões ativas. (Le Roux; Rouanet, 2010, p. 118). Esse é o caso especialmente quando as categorias são inseridas como passivas por indicarem **respostas ausentes**.[28] Tais indivíduos podem ser inseridos como passivos, ou seja, suas "escolhas" nas modalidades de resposta não serão consideradas para a construção dos eixos e para a definição das distâncias relativas.

De forma similar ao exemplo anteriormente apresentado, há 33 variáveis ativas e 72 modalidades ativas, mais 33 modalidades passivas. O número de eixos aqui é calculado de forma ligeiramente diferente, conforme a seguinte fórmula: $K' - (Q - Q')$, em que K' se refere ao número de categorias ativas; e Q', ao número de questões com, ao menos, uma categoria passiva. Como todas as variáveis têm, ao menos, uma categoria desse tipo, então o número total de eixos equivale ao número de categorias ativas, ou seja, 72.

Como podemos observar na Tabela 1.5, a seguir, os resultados da ACM Específica são bastante similares aos da ACM padrão, ainda que a

[27] É importante ressaltar que as categorias de não resposta podem ser pertinentes sociologicamente, dependendo do objeto de investigação. Por exemplo, caso se esteja investigando tomadas de posição no plano político, a distribuição das não respostas pode ser um indicador muito importante da desigual capacidade dos agentes de produzir opiniões sobre questões políticas ou mesmo de apreender a problemática política de um tema qualquer (Bourdieu, 1984).

[28] No exemplo que mostraremos a seguir, 16 indivíduos serão inseridos como passivos.

quantidade de casos seja diferente.[29] Considerando os critérios para a escolha do número de eixos a reter, chega-se, igualmente, a três eixos. A interpretação desses eixos revela contrastes ou oposições idênticas às já descritas.[30]

Tabela 1.5 | Autovalores e proporções da inércia dos primeiros eixos: comparação entre ACM convencional e ACM específica

Eixo	Autovalor (ACM padrão)	Autovalor (ACM específica)	% da variância (ACM padrão)	% da variância (ACM específica)	% variância modificada (ACM padrão)	% variância modificada (ACM específica)
1	0,1665	0,177	14,09	14,9	69,98	64,1
2	0,1116	0,113	9,44	9,4	24,91	22,7
3	0,0596	0,059	5,04	5,0	3,23	4,3
4	0,0457	0,045	3,87	3,8	0,89	1,9
5	0,0419	0,040	3,55	3,4	0,51	1,2

Fonte: elaboração própria.

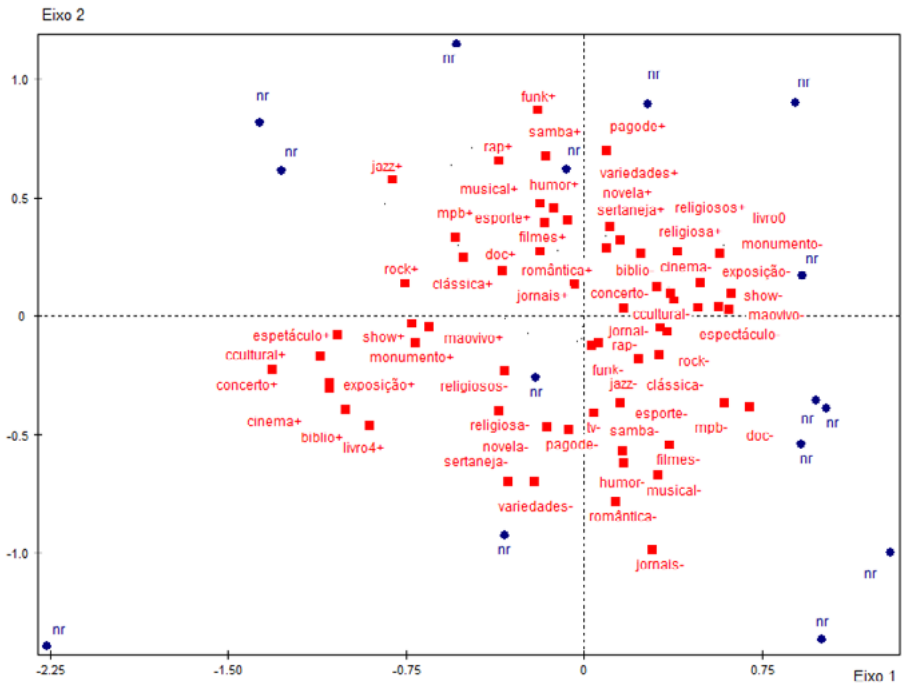
As categorias passivas, ainda que "desmagnetizadas" e, portanto, sem qualquer contribuição para a construção das distâncias relativas, podem ser projetadas na nuvem de modalidades, assim como os indivíduos passivos, cujas respostas são igualmente descartadas para a determinação dos eixos. A seguir, exibimos as nuvens de indivíduos e de modalidades, formadas pelos eixos 1 e 2, diferenciando as categorias ativas das passivas e os indivíduos ativos dos passivos. No caso da nuvem de modalidades, notemos que as categorias de não resposta (representadas por círculos) estão localizadas relativamente longe do centro do plano cartesiano, pois são geralmente pouco frequentes.

[29] Vale notar que essa não é a regra. A base de dados utilizada no exemplo anterior (da ACM convencional) excluía os indivíduos com respostas ausentes. Este exemplo utiliza a base de dados com alguns indivíduos que não forneceram resposta a uma ou mais questões. Como as modalidades excluídas são apenas as que indicam resposta ausente (e, para os propósitos deste exercício, as não respostas não têm qualquer pertinência empírica), os resultados são bem parecidos. A base de dados da ACM específica possui 1122 casos; aquela da ACM convencional, 995 casos. Aplicando as ponderações pertinentes, esses valores são, respectivamente: 1122 e 974.

[30] Por isso, os resultados não serão apresentados detalhadamente.

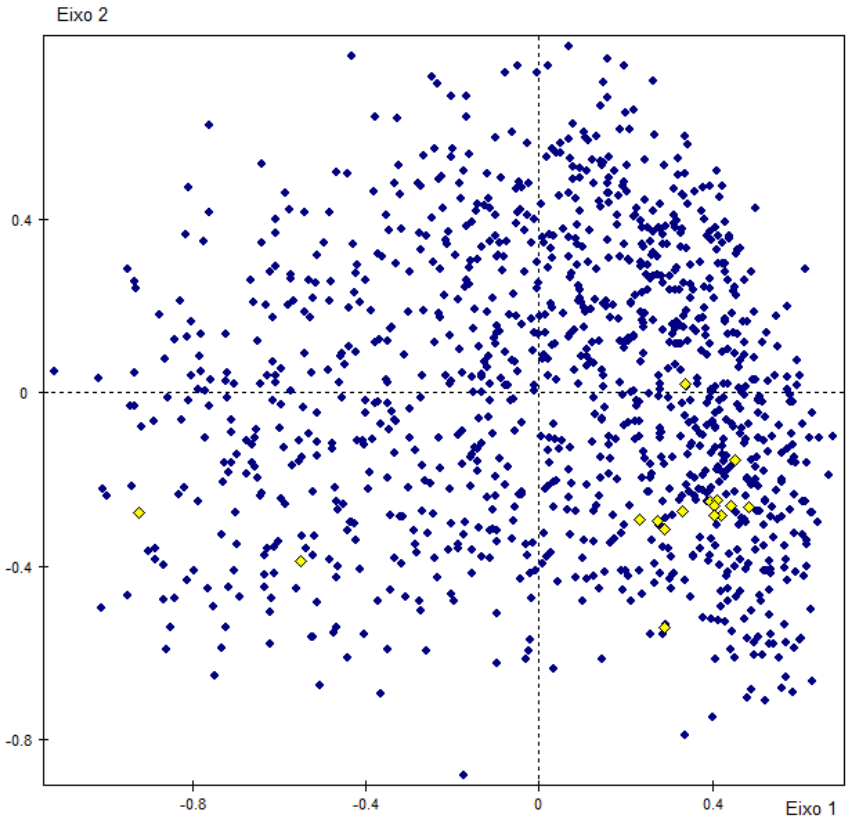
Isso significa que, mesmo que as considerássemos sociologicamente pertinentes, elas deveriam ser inseridas como passivas, para minimizar esse efeito de "esticamento" da nuvem com a consequente concentração de muitas categorias ao redor da origem do espaço.

Figura 1.7 | Categorias ativas (quadrados) e passivas (círculos) projetadas no plano fatorial formado pelos eixos 1 e 2 (ACM específica)



Fonte: elaboração própria.

Figura 1.8 | Nuvem de indivíduos ativos (círculos) e passivos (losangos) no plano fatorial formado pelos eixos 1 e 2 (ACM específica)

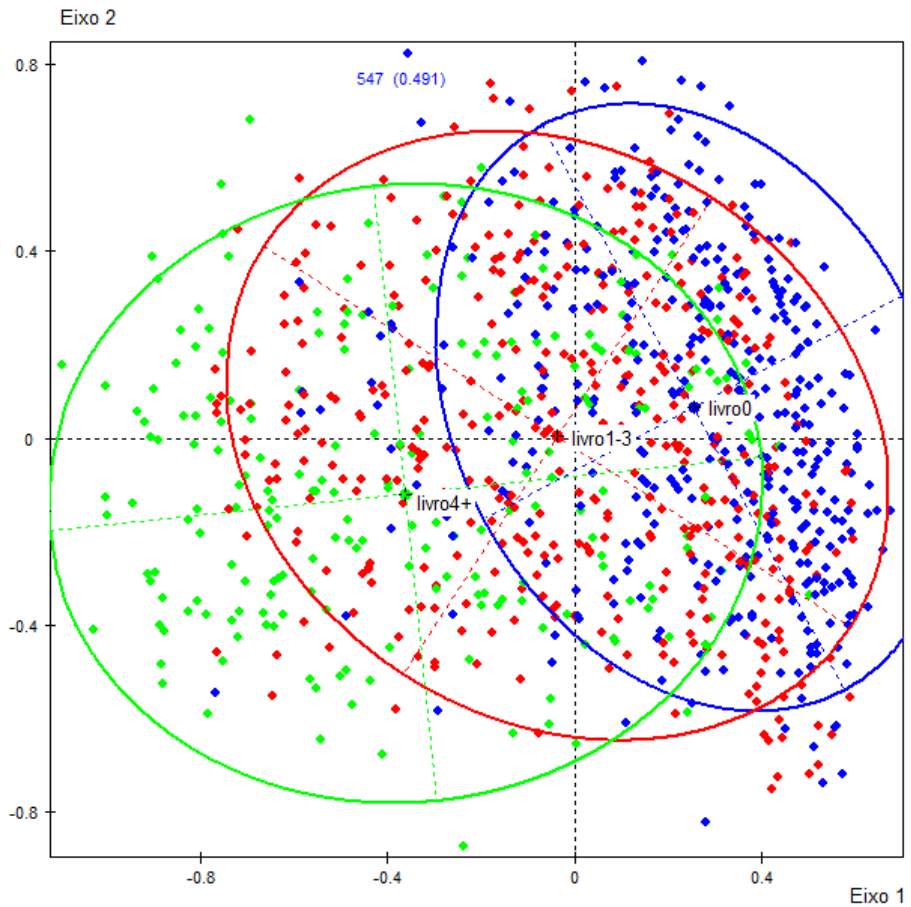


Fonte: elaboração própria.

A nuvem de indivíduos: primeiras inspeções

As nuvens de modalidades e de indivíduos possuem a mesma dimensionalidade, em termos do número e da orientação dos eixos e do peso de cada um para a inércia total. Por meio das chamadas **fórmulas de transição** (Le Roux; Rouanet, 2004, p. 41-2), é possível passar de uma nuvem para outra. Tais fórmulas permitem calcular, por exemplo, a coordenada de um indivíduo em determinado eixo a partir de seu padrão de resposta. Observando a Figura 1.9, vemos que o ponto que representa o indivíduo 547 está à esquerda da origem do eixo 1 (-0,35) e acima da origem no eixo 2 (0,82). Seu padrão de resposta é relativamente atípico; daí estar distante do centro do plano fatorial ou do baricentro. Este representa uma espécie de "perfil médio". Como argumenta Duval (2015), os indivíduos que, para o conjunto das variáveis contidas na análise, são os mais "atípicos" em relação ao "perfil médio" contribuem muito na configuração da nuvem e, por conseguinte, na construção dos eixos. Sabendo o padrão de resposta de um indivíduo e conhecendo as coordenadas fatoriais das categorias "escolhidas", então é possível calcular a localização do indivíduo em determinado eixo, conforme a fórmula a seguir: $y^i = \frac{1}{\sqrt{\lambda}} \sum_{k \in K_i} \frac{y^k}{Q}$, ou seja, a coordenada de um indivíduo em um eixo é o produto da soma das coordenadas fatoriais das categorias "escolhidas" (y^k) dividida pelo número de questões (Q), valor, por sua vez, dividido pela raiz quadrada da inércia (λ) do respectivo eixo.

Figura 1.9 | Elipses de concentração e subnuvens resultantes da partição da nuvem global de indivíduos conforme as categorias da variável "leitura de livros"



Fonte: elaboração própria.

Se partirmos da nuvem de indivíduos, é possível nela plotar os **pontos médios** das categorias ativas e suplementares. Por definição, o ponto médio de uma categoria é o "ponto médio da subnuvem de indivíduos que escolheram essa categoria" (Le Roux; Rouanet, 2004, p. 42). Cada questão *q* permite particionar a nuvem de indivíduos conforme o número de modalidades K_q. Por exemplo, a variável "livros" possui três categorias

“livro0”, “livro1-3” e “livro4+”, o que nos permite criar, portanto, três subnuvens de indivíduos, conforme se vê na Figura 1.9.

Inicialmente, podemos investigar a dispersão dos indivíduos ao redor dos pontos médios de diferentes categorias nos diversos planos fatoriais construídos a partir dos eixos retidos para interpretação. A principal ferramenta para essa investigação é a **elipse de concentração**, que concentra o equivalente a 86,5% dos casos em uma distribuição bidimensional ou 95,4% deles em uma distribuição unidimensional. Observando a Figura 1.9, percebe-se que as elipses se distribuem claramente ao longo do primeiro eixo, estando aquela que concentra os indivíduos que “escolheram” a modalidade “livro4+” mais à esquerda e aquela dos indivíduos que “escolheram” a categoria “livro0”, mais à direita. Nota-se, também, que as elipses possuem formatos diferentes, o que nos dá alguns indícios das diferenças entre elas em termos da variância ao longo dos dois eixos que constituem o plano fatorial: a elipse verde se “espalha” ao longo do primeiro eixo, ocupando uma porção maior dele que as outras, o que significa que os indivíduos que “escolheram” a categoria representada por essa elipse diferem bastante – ou, pelo menos, em maior medida – em termos dos contrastes desse eixo principal. Diferentemente, a elipse azul possui um formato mais ovalado, ocupando uma porção do primeiro eixo quase totalmente circunscrita a um de seus lados, o que significa que essa subnuvem de indivíduos é mais homogênea no que se refere a tais contrastes. Em outras palavras, não ler nenhum livro é um indicador razoavelmente preciso do (baixo) engajamento cultural do indivíduo, enquanto ler muitos livros (no caso, 4 ou mais nos últimos seis meses) não constitui um indicador preciso de elevado engajamento cultural, pelo menos no que se refere à frequência das “saídas culturais”.

Conclusões

O princípio que orienta a ACM (e, mais amplamente, a AGD) é o de que "o modelo deve seguir os dados, e não o contrário". Diferentemente de outras técnicas estatísticas voltadas para a busca do modelo que mais se ajusta aos dados, a ACM é uma técnica que permite apreender, por meio da análise visual de planos fatoriais, **a estrutura de relações entre conjuntos de modalidades e de agentes**. Como argumenta Di Franco (2016, p. 1304), "o propósito da ACM é criar apenas algumas dimensões capazes de reproduzir a maior parte da inércia entre as variáveis categóricas analisadas em um pequeno número de fatores expressando combinações de todas as categorias ativas". A interpretação das nuvens é relativamente direta: quanto maior a proximidade de duas categorias, mais estão diretamente associadas; quanto mais próximos dois indivíduos estiverem, mais similares são seus perfis de respostas; quanto mais distante uma categoria estiver do centro do eixo, mais ela contribuirá para a formação daquele eixo (e, portanto, para sua inércia).

É importante notar, como já demonstrado, que a ACM (bem como outras variantes da AGD) possibilita analisar **simultaneamente as relações entre categorias e indivíduos**: por meio das fórmulas de transição, como vimos, é possível passar da nuvem de indivíduos para a nuvem de categorias, projetando os indivíduos na nuvem de categorias, assim como as categorias na nuvem de indivíduos.

Entre as críticas à ACM, Di Franco (2016) sublinha três principais:

> i. essa técnica teria um objetivo meramente descritivo, produzindo resultados que poderiam ser visualizados de outros modos mais simples (por exemplo, por meio da leitura de um conjunto de tabelas de contingência ou descritivas);
>
> ii. os resultados produzidos pela ACM seriam muito instáveis, dependendo da amostra utilizada, das definições operacionais e das escolhas quanto a quais variáveis incluir como ativas;
>
> iii. por não se basear em estatísticas inferenciais, os resultados não seriam generalizáveis.

Em relação à primeira crítica, Di Franco (2016) argumenta que o que se entende por limitação é, ao invés, a principal vantagem da ACM: a possibilidade de analisar as relações entre um grande conjunto de variáveis **e** de indivíduos, que não poderiam ser apreendidas por meio da leitura de um conjunto de tabelas, nem mesmo de um grande número delas, porque não teríamos a possibilidade de analisar tais relações **simultaneamente**. Ou seja, a ACM possibilita apreender a **estrutura de relações** ou de **afinidades** e **contrastes** entre as categorias e os indivíduos e é somente nessa estrutura de relações que a localização de uma categoria ou de um indivíduo ganha sentido.

Em relação à segunda crítica, o autor sugere que sejam utilizadas amostras relativamente grandes (com, no mínimo, 20 casos para cada categoria ativa utilizada); que se atente para os casos anômalos ou as categorias pouco frequentes (inserindo-os como passivos), e que se balanceie o número de categorias de variáveis (e, como já mencionado, de variáveis por tópico).

Em relação à terceira crítica, é importante ressaltar que há modos de verificar a estabilidade das soluções fatoriais e de calcular intervalos de confiança para os principais parâmetros da ACM. Além disso, como sugerem Le Roux e Rouanet (2010, p. 299), a distinção entre procedimentos descritivos e inferenciais é apenas aparente: na verdade, "todos os métodos estatísticos produzem estatísticas básicas que são descritivas... que, em um estágio mais avançado, podem ser combinadas com o tamanho da amostra para produzir procedimentos indutivos". Nesse sentido, a orientação descritivo-exploratória da ACM traria uma vantagem em relação a técnicas inferenciais: como a descrição vem em primeiro lugar e, apenas depois, a análise inicial é corroborada por procedimentos que buscam verificar a estabilidade dos resultados, não se gasta tempo demais com a análise de resultados que, embora possam ser estatisticamente significativos, são negligenciáveis do ponto de vista descritivo.[31]

[31] Cf. Le Roux e Rouanet (2010, p. 297-332) e Di Franco (2016, p. 1306-1307).

Capítulo 2

Análise de dados estruturados: construindo e interpretando a nuvem de indivíduos

A definição das variáveis e categorias como ativas ou suplementares depende essencialmente das questões de pesquisa. Gostaria de apresentar brevemente dois estudos para exemplificar esse argumento.

Em *Culture, class, distinction* (Bennett *et al.*, 2009), os principais objetivos da pesquisa consistiam em: i) examinar os processos de (re)produção e transmissão de capital cultural na sociedade britânica contemporânea e a forma que tais processos assumiam; e ii) investigar se havia alguma homologia, em termos dos princípios de estruturação das práticas, em diferentes campos culturais (música, televisão, leitura, arte, filmes, esportes, alimentação

> fora). Para isso, então, os autores construíram "mapas culturais" inserindo como **variáveis/modalidades ativas** as questões sobre **práticas**, **gostos** e **conhecimentos** nesses sete domínios culturais, de modo a apreender as "relações mútuas entre aspectos da vida cultural em si". Uma vez construídos os mapas, foram sobrepostas as **variáveis suplementares** (especialmente idade, sexo, classe ocupacional, escolaridade e renda), para determinar se tais variáveis estavam "associadas com a paisagem cultural" (Bennett *et al*., 2009, p. 44).

Partindo de uma problemática distinta – ainda que articulando perspectivas teóricas similares –, o sociólogo português José Virgílio Borges Pereira, em seu estudo *Classes e culturas de classe das famílias portuenses*, buscou, em um primeiro momento, reconstruir o espaço social citadino, em termos da configuração e do relevo "dos principais processos de divisão social", com ênfase no "impacto dos processos de diferenciação social de tipo classista". Num segundo momento, o estudo tentou equacionar "a leitura e compreensão das práticas quotidianas dos agentes da cidade e dos protagonismos que estão na base de sua produção". Ora, partiu-se, assim, da noção de espaço social – um sistema multidimensional de coordenadas definidas em função da distribuição dos diferentes tipos de capital entre os agentes –, como estruturante das práticas e representações por intermédio dos esquemas de percepção e de avaliação (*habitus*). O objetivo era examinar as homologias entre os espaços social e simbólico. Do ponto de vista metodológico, o estudo fez amplo uso da ACM, utilizando como variáveis/modalidades ativas indicadores de capital econômico (tipo de moradia, pertencimento de classe), cultural (escolaridade, número de livros possuídos) e social (número de contatos na agenda telefônica) para a construção do espaço social (Pereira, 2005, p. 190-218). Feito isso, as modalidades referentes às práticas de usos do tempo livre foram projetadas como **suplementares** de forma a caracterizar as diferentes regiões (ou "zonas", como prefere o autor) do espaço social em termos de conjuntos sistemáticos de práticas sociais e culturais.

O que os dois exemplos anteriores nos mostram é que, se quisermos examinar as relações entre dois ou mais conjuntos de variáveis (que, na verdade, nos permitem classificar os agentes com base em indicadores de diferentes conjuntos de propriedades pertinentes à análise), uma possibilidade é definir um conjunto como **ativo** e o(s) outro(s) como **suplementar(es).** Considerando três conjuntos principais de variáveis – as **sociodemográficas** (sexo, idade, escolaridade, renda etc.), aquelas sobre **práticas/comportamentos** (frequência com que alguém vai ao cinema, faz viagens internacionais, se conhece outros idiomas etc.) e, por fim, as **atitudinais/perceptivas** (opinião sobre o que é mais importante fazer quando se viaja para fora ou sobre a arte moderna e outros estilos artísticos) –, se inserirmos as variáveis de um conjunto como ativas, então as demais deverão ser projetadas posteriormente como suplementares.

Conforme argumenta Hjellbrekke (2019, p. 64), "[a] inclusão de variáveis suplementares abre a possibilidade de examinar a nuvem de indivíduos ao incluir fatores estruturantes na análise. Fatores estruturantes são variáveis ou descritores que não tomam parte na definição das distâncias nas nuvens". Trata-se do que Le Roux e Rouanet (2004, 2010) designam por **análise de dados estruturados** [*structured data analysis*] – bases de dados equipadas com "fatores estruturantes" –, que possibilita integrar as técnicas usadas para manejar tais fatores (como a análise de variância) no âmbito da AGD, preservando suas principais características (Le Roux; Rouanet, 2010, p. 68). Outra maneira de caracterizar a análise de dados estruturados é compará-la ao uso de regressões na análise multivariada "convencional", pois ela também possui uma orientação **preditiva**: sabendo o valor de um indivíduo em dado fator estruturante (seu sexo ou sua idade), onde esse indivíduo provavelmente estará localizado no plano fatorial? Ou, então, sabendo onde o indivíduo está localizado nesse espaço, qual é seu provável "valor" nesse fator estruturante?

Le Roux e Rouanet (2004) referem-se neste ponto ao "uso explicativo" da AGD. Nos estudos de classe de Bourdieu, conforme o exemplo que eles usam, há duas questões importantes: i) como as posições relativas no espaço social são explicadas pelas frações de classe a que os agentes

pertencem?; ii) como as posições no espaço social explicam as tomadas de posição dos agentes em diversos domínios?

Essa propriedade da análise de dados estruturados – e, mais amplamente, da simetria das nuvens de indivíduos e de modalidades – pode ser bastante útil para a tarefa de construção de amostras qualitativas em pesquisas multimétodos. Por exemplo, é possível escolher indivíduos que são os portadores típicos de um conjunto minimamente coerente de gostos, práticas e representações em dada região do espaço (por exemplo, o estilo de vida operário) ou, então, ao revés, escolher os casos mais improváveis.[1]

A análise de dados estruturados se dá em três passos:

i. O estudo das distâncias entre as **coordenadas fatoriais** das categorias suplementares projetadas na nuvem de categorias. A fórmula para o cálculo da coordenada y^k do ponto de uma categoria M^k é dada por: $y^k = \frac{1}{\sqrt{\lambda}} \sum_{i \in I_k} \frac{y^i}{n_k}$, ou seja, é a simples média das coordenadas y^i dos n^k pontos de indivíduos que escolheram uma categoria k, dividida por $\sqrt{\lambda}$. Esta é a segunda **fórmula de transição,** que permite passar da nuvem de indivíduos para a nuvem de categorias (Le Roux; Rouanet, 2010, p. 41).[2]

ii. O exame da **dispersão dos indivíduos em uma subnuvem** por meio da construção das elipses de concentração. Aqui também outras propriedades são observadas: a **excentricidade** das elipses e a **decomposição da variância das subnuvens,** resultantes da partição da nuvem de indivíduos conforme as categorias de uma variável suplementar, em termos das variâncias **internuvem** [*between-variance*] e **intranuvem** [*within-variance*].

iii. A análise dos efeitos do cruzamento entre dois fatores estruturantes, diferenciando entre os efeitos principais [*main effects*] e os efeitos **dentro** das subcategorias [*within effects*].

[1] Cf. Bennett *et al*. (2019, p. 24-39).

[2] Ou seja, essa fórmula permite calcularmos a coordenada fatorial de uma categoria na nuvem de modalidades a partir do ponto médio da modalidade na nuvem de indivíduos. Como sublinham Le Roux e Rouanet (2010, p. 234), "a diferença das coordenadas entre as modalidades ao longo de um eixo na nuvem de modalidades é igual ao desvio entre os correspondentes pontos médios das modalidades na nuvem de indivíduos expresso em unidades de desvio-padrão". Hjellbrekke (2019) propõe a análise dos desvios entre os pontos médios das modalidades, mas a simples comparação da diferença entre as coordenadas fatoriais na nuvem de categorias produz os mesmos resultados.

Inicialmente é importante fazer uma análise preliminar da associação entre os dois conjuntos de variáveis, as ativas e as suplementares, por meio da leitura das coordenadas fatoriais das modalidades suplementares, com ênfase nos eixos que foram retidos para interpretação.

A projeção das categorias das variáveis suplementares no espaço fatorial pode produzir três resultados típicos:

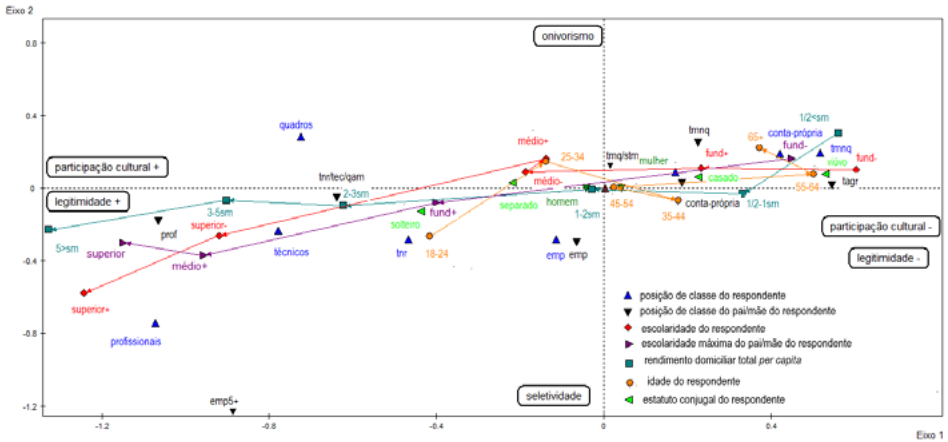
i. **forte concentração em torno do centro do plano fatorial**, o que indica haver pouca ou nenhuma associação entre os dois conjuntos de variáveis (ativas e suplementares);

ii. **forte dispersão em torno do centro do plano fatorial**, indicando provavelmente uma forte associação entre esses dois conjuntos;

iii. **combinação entre concentração e dispersão**, o que é, na verdade, o resultado mais comum, indicando forte associação para algumas variáveis e pouca ou nenhuma, para outras.

Nas análises que usamos como exemplos práticos da aplicação da ACM, buscou-se apreender as relações entre diferentes aspectos das práticas e gostos culturais, cujas variáveis foram inseridas como ativas. Para isso, foram construídos três "mapas culturais", acompanhados das descrições dos principais contrastes observados nos respectivos eixos. A base de dados utilizada para tais exercícios traz também variáveis sociodemográficas. O objetivo, então, é saber **se** e **em que medida** tais variáveis estão relacionadas com os eixos retidos para interpretação no momento anterior. Estes são os **fatores estruturantes**, ou seja, os "descritores" que não servem para a definição das distâncias relativas.

Figura 2.1 | Projeção das categorias suplementares no plano fatorial formado pelos eixos 1 e 2



Fonte: elaboração própria.

Observando a Figura 2.1, vemos que as categorias de algumas variáveis suplementares se "espalham" ao longo do primeiro eixo (renda, escolaridade do respondente, idade, escolaridade máxima dos pais, classe do respondente e origem de classe). No caso das variáveis ordinais, as **setas** descrevendo as **trajetórias** conectando as categorias em ordem ascendente são bastante úteis para evidenciar o padrão da distribuição no plano fatorial. Por exemplo, as categorias das variáveis **escolaridade** (do respondente e a mais elevada do pai ou da mãe) e **renda** se distribuem do menor ao maior valor à medida que passamos da direita para a esquerda do primeiro eixo.

Ainda que em um padrão menos regular, as categorias da variável idade se distribuem ao longo do primeiro eixo de forma similar às anteriores: as categorias dos mais jovens estão à esquerda ("18-24" e "25-34"); enquanto as dos mais idosos ("55-64" e "65+"), à direita; os pontos das outras duas categorias dessa variável ("35-44" e "45-54") estão também à direita, embora relativamente mais próximos do centro do plano fatorial. Podemos calcular as localizações no plano fatorial das categorias dos mais jovens (entre 18 e 34 anos) e dos mais idosos (55 ou mais), usando

a seguinte fórmula: $\frac{(y_k \cdot n_k) + (y_{k'} \cdot n_{k'})}{n_k + n_{k'}}$. Disso resulta que a coordenada da categoria dos mais jovens ("18-34") no eixo 1 é **-0,25**; e aquela da categoria dos mais idosos ("55+"), **0,44**.

É possível também observar padrões na distribuição das variáveis de classe (do respondente e a do pai/mãe) ao longo do referido eixo[3]: as categorias não manuais (profissionais, técnicos, quadros e trabalhadores não manuais de rotina) estão à esquerda do eixo, enquanto aquelas das categorias de trabalhadores manuais estão à direita do eixo horizontal.

As categorias da variável **estatuto conjugal** também se espalham ao longo do primeiro eixo: "solteiro" e "separado" estão à direita; "casado" e "viúvo", à esquerda. Notemos, por fim, que as categorias da variável **sexo** estão bem próximas entre si, no centro do plano fatorial.

O que podemos apreender dessa inspeção inicial das localizações das modalidades suplementares no plano fatorial formado pelos eixos 1 e 2? Há, como notamos, **um padrão que combina dispersão e concentração**: as categorias das variáveis escolaridade, renda e classe estão fortemente dispersas; aquelas das variáveis idade e estatuto conjugal também, mas em menor medida; e, por fim, as categorias da variável sexo estão concentradas no centro do plano fatorial. A dispersão e concentração são indícios da correlação de uma variável com o eixo em questão. Tal análise, como mencionado, tem uma **orientação preditiva**, ou seja, sabendo, por exemplo, que um indivíduo tem ensino superior ou é um profissional ou tem origem em uma família em que o pai e/ou a mãe possuía ensino superior, temos boas chances de inferir o padrão de suas respostas, caracterizado por elevada participação cultural e consumo de bens culturais "legítimos". Ao invés, sabendo que outro indivíduo é trabalhador manual, tem origem em meios rurais ou urbano-manuais e possui baixa escolaridade, é provável que seu padrão de respostas seja caracterizado por baixo participação cultural e ausência de consumo dos bens da "alta cultura".

[3] As modalidades referentes à posição de classe do respondente são representadas por triângulos; aquelas referentes à posição de classe de origem, por triângulos invertidos.

Essa inspeção inicial, a partir da observação da nuvem de categorias suplementares, deve ser complementada (e eventualmente corrigida) por **uma análise das distâncias entre as coordenadas fatoriais das categorias de uma mesma variável**. Distâncias maiores do que 0,5 são consideradas **notáveis** e aquelas maiores do que 1 são considerados **grandes** (Le Roux; Rouanet, 2010, p. 59). Considerando o eixo horizontal (1), temos que a distância entre as categorias da variável "idade", que inclui os mais jovens (entre 18 e 34 anos), de um lado, e os mais velhos (acima de 55 anos), de outro [(0,44) – (-0,25) = **0,69**], pode ser considerada notável ao longo do eixo horizontal, e grande entre aquelas que indicam que o respondente possui ensino superior completo ("superior+") ou ensino médio completo ("médio+") [1,10], assim como entre ensino superior completo ("superior+") e fundamental incompleto ("fund-") [1,84]. No mesmo eixo, a distância é notável entre as modalidades "profissionais" (profissionais) e trabalhadores não manuais de rotina ("tnr") [0,6] e grande entre "profissionais" e trabalhadores manuais qualificados e supervisores de trabalho manual ("tmq/stm") [1,59]. Já entre "homem" e "mulher", a distância é de 0,08, ou seja, não significativa.

Tanto a inspeção inicial da nuvem de categorias suplementares quanto o cálculo dos desvios nos permitem afirmar que algumas variáveis estão correlacionadas com o eixo horizontal (1) e que sexo não tem qualquer correlação com ele. Desse ponto de vista, saber se um indivíduo é do sexo masculino ou feminino pouco nos ajuda a prever sua localização no plano fatorial.

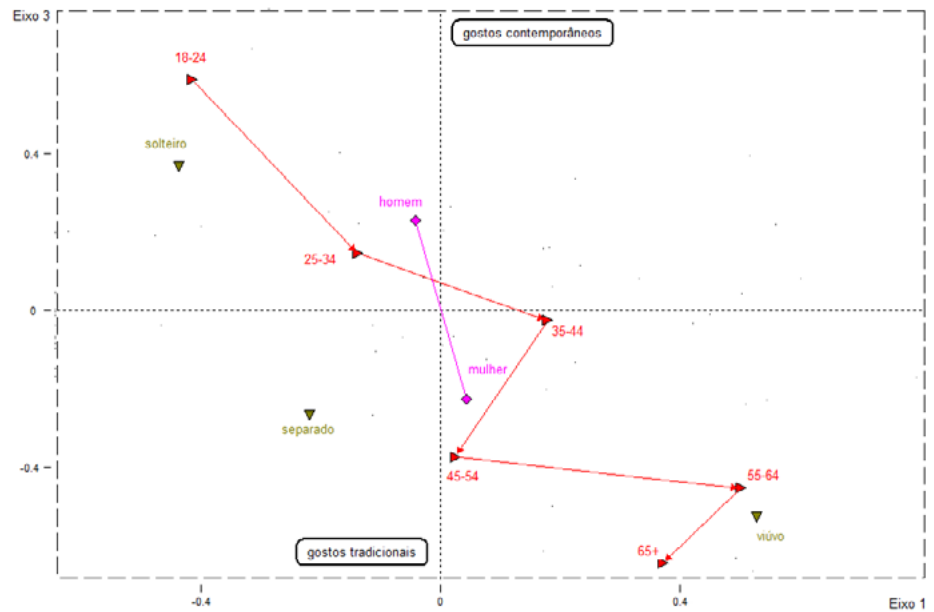
Considerando agora a distribuição das categorias suplementares pelo eixo vertical (2), conforme a Figura 2.1, há padrões menos definidos: as categorias "profissional" (agregado ocupacional do respondente) e "emp5+" (agregado ocupacional de origem) são as que estão relativamente mais distantes da origem do segundo eixo (na região inferior), em comparação com as categorias "quadros", "tmnq" e "cp" (agregado ocupacional do respondente) e "tmnq" (agregado ocupacional de origem), que ocupam posições opostas, embora relativamente mais próximas à origem. As distâncias da categoria "profissional" em relação às outras, acima do eixo, são, respectivamente: 1,02 (grande); 0,93 (grande); 0,93 (grande); e aquele da categoria "emp5+" em relação à categoria

da mesma variável (“tmnq”) é 1,48 (grande). Ou seja, considerando as oposições do segundo eixo (presença *versus* ausência de gostos “onívoros”), vemos que, entre os profissionais com origem em famílias em que o pai e/ou a mãe também eram profissionais ou empregadores, são maiores as chances de rejeição a um repertório de gosto que combina gêneros culturais muito diversos entre si. Diferentemente, sabendo que um indivíduo é um trabalhador manual ou quadro gerencial, com origem em meios populares, é mais provável que manifeste repertórios de gosto (em termos de música e de programas televisivos) “onívoros”.

Por fim, no que se refere ao terceiro eixo (Figura 2.2), a oposição nele apreendida, entre aderência à “cultura clássica” *versus* aderência à “cultura contemporânea”, está correlacionada com idade e com estatuto conjugal. Em relação à primeira, observando as setas que formam a trajetória conectando as diferentes categorias de idade, vemos que as faixas etárias se distribuem, da menor à maior, à medida que descemos o eixo vertical (3). Isso significa que, enquanto os mais jovens tendem a possuir repertórios de gosto mais contemporâneos, os mais idosos, por sua vez, têm maior protagonismo nos repertórios “clássicos” ou “tradicionais”. Argumento similar vale para solteiros, na parte de cima do eixo, em comparação com os “separados” e “viúvos”, na parte de baixo. Diferentemente do que observado no exame dos outros eixos, neste é possível perceber que as categorias da variável sexo estão relativamente distantes entre si. Uma inspeção visual nos levaria a crer que tal eixo estaria correlacionado com sexo. No entanto, a distância entre as coordenadas de tais categorias é de apenas 0,46, portanto, abaixo, embora bastante próximo, do valor que poderíamos considerar como notável.

Podemos, então, descartar o argumento de que o eixo 3 está também estruturado por sexo? A distância entre as categorias da variável sexo pode ser considerada significativa, ainda que menor que o valor limite estabelecido pelo critério anterior? Voltaremos a essas questões adiante.

Figura 2.2 | Projeção das modalidades suplementares no plano fatorial formado pelos eixos 1 e 3



Fonte: elaboração própria.

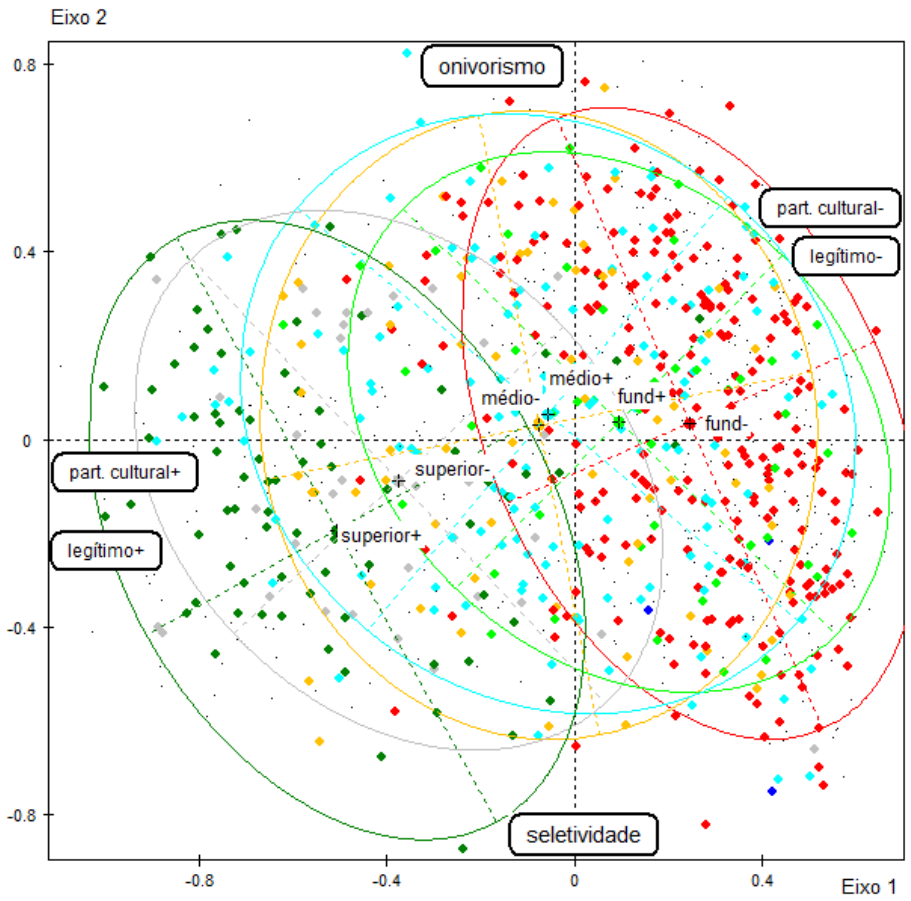
A inspeção visual e o cálculo das distâncias entre as categorias de uma mesma variável ao longo de um ou mais eixos constituem **passos iniciais** da análise de dados estruturados. Se interrompêssemos a análise aqui, ficaríamos restritos à comparação das coordenadas fatoriais das categorias suplementares, deixando de lado outras informações tão ou mais importantes, como a dispersão dos indivíduos caracterizados por uma categoria no plano fatorial. Por isso, os passos seguintes incluem o exame da **dispersão dos pontos** (que representam os indivíduos) ao redor do **ponto médio de uma subnuvem** (por meio das **elipses de concentração**) e a **desagregação da variância da nuvem global** em termos da variância **entre** subnuvens e da variância **dentro** das subnuvens, que evidenciam quão homogêneas são as referidas subnuvens resultantes da partição da

nuvem global de indivíduos e quão fortemente diferem entre si ao longo de um eixo (ou plano fatorial).[4]

Uma elipse de concentração inclui aproximadamente 86,5% dos pontos de uma subnuvem em uma distribuição bidimensional e 95,4% em uma distribuição unidimensional. Ao examiná-la, devemos estar atentos para seu formato (se esférica ou ovalada) e, correlatamente, para o modo como se "esparrama" em um plano fatorial. O grau de excentricidade, que varia de 0 a 1, de uma elipse de concentração nos dá uma informação importante: quanto mais próximo de 0, mais a elipse apresenta um formato esférico; quanto mais próximo de 1, mais seu formato será ovalado. **O formato nos dá indícios importantes sobre o grau de homogeneidade dos pontos de uma subnuvem nos eixos que constituem um plano fatorial**. Uma elipse ovalada, por exemplo, indica maior heterogeneidade ao longo de um eixo e, ao revés, maior homogeneidade ao longo do outro. Tomemos, como exemplos, as elipses de concentração das subnuvens da variável "escolaridade" no plano fatorial formado pelos eixos 1 e 2.

[4] O termo ponto médio [*mean point*] de uma categoria se refere ao ponto médio da subnuvem de indivíduos que "escolheu" uma categoria. Por meio das fórmulas de transição, como mostrei anteriormente, é possível calcular a coordenada fatorial de uma categoria na nuvem de modalidades a partir de seu ponto médio na nuvem de indivíduos (Le Roux; Rouanet, 2010, p. 42). Por isso, é importante reter essa importante diferenciação entre ponto médio $(\bar{y})$ de uma subnuvem na nuvem de indivíduos e a coordenada fatorial (y) de uma categoria na nuvem de modalidades.

Figura 2.3 | Elipses de concentração e subnuvens resultantes da partição da nuvem global de indivíduos conforme as categorias da variável "escolaridade" (eixos 1 x 2)



Fonte: elaboração própria.

Vemos, na Figura 2.3, que há alguma sobreposição entre as elipses, embora não completa, o que é um resultado bastante comum quando um eixo é estruturado pela variável cujas categorias serviram para particionar a nuvem global. Não houvesse qualquer correlação entre o eixo e a variável, as elipses de concentração das subnuvens se sobreporiam quase completamente. Vemos, ainda, que as elipses se "esparramam"

desde o quadrante superior à direita (elipse em vermelho engloba 88,8% dos pontos da subnuvem formada pelos menos escolarizados) até o quadrante inferior à esquerda (elipse que inclui 89,6% dos pontos da subnuvem formada pelos indivíduos mais escolarizados), o que significa que a variável escolaridade apresenta **uma correlação com os dois eixos que formam o referido plano fatorial** (ainda que mais fortemente com o eixo 1). A elipse da subnuvem formada pela categoria "fund-" (ensino fundamental incompleto) ocupa a região do quadrante "participação cultural-/legitimidade-/onivorismo", enquanto aquela da categoria "superior+", por sua vez, ocupa a região do quadrante "participação cultural+/legitimidade+/seletividade". As demais elipses ocupam porções intermediárias do plano fatorial entre esses dois quadrantes.

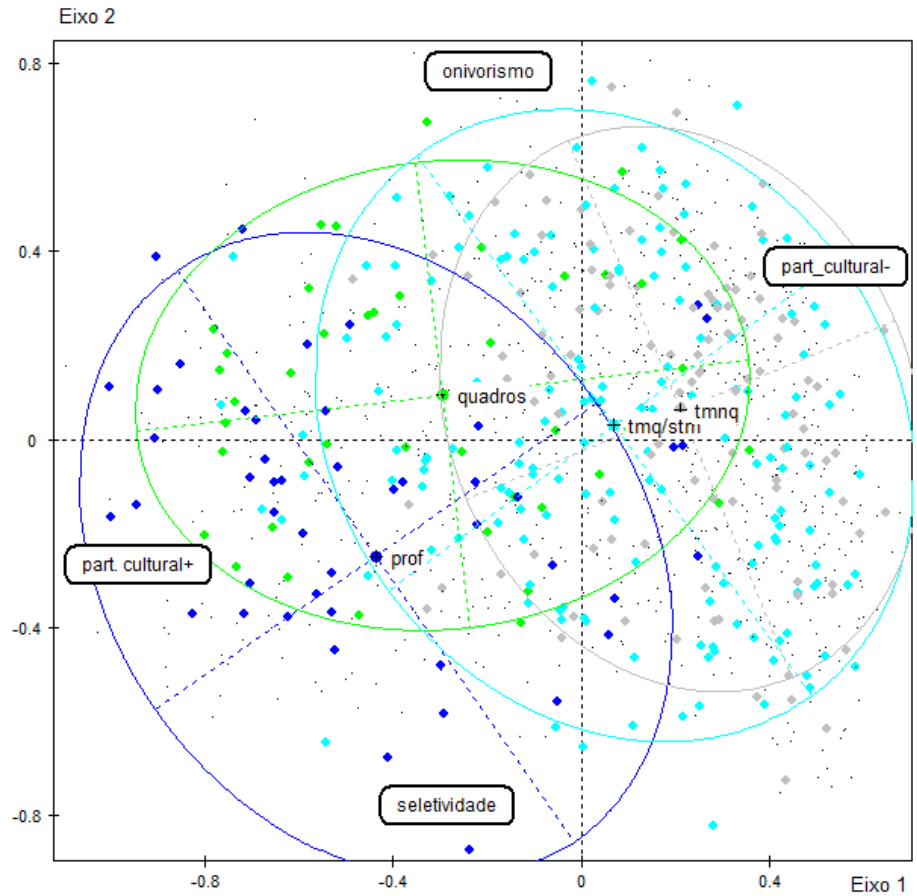
Tabela 2.1 | Excentricidade das elipses de concentração, proporção de pontos dentro das elipses e coordenadas dos pontos médios das subnuvens

Categorias	Excentricidade	Proporção de pontos dentro da elipse (%)	Coordenadas dos pontos médios	
			Eixo 1	Eixo 2
superior+ (ensino superior completo)	0,759	89,6	-0,507	-0,193
superior- (ensino superior incompleto)	0,673	86,1	-0,375	-0,087
médio+ (ensino médio completo)	0,497	84,8	-0,056	0,055
médio- (ensino médio incompleto)	0,479	89,2	-0,076	0,031
fund+ (ensino fundamental completo)	0,642	90,9	0,095	0,037
fund- (ensino fundamental incompleto)	0,796	88,8	0,246	0,035
prof (profissionais)	0,638	85,4	-0,437	-0,248
quadros (diretores, gerentes)	0,651	88,1	-0,296	0,095
tmq/stm (trabalhadores manuais qualificados)	0,551	91,2	0,07	0,03
tmnq (trabalhadores manuais não qualificados)	0,604	86,8	0,211	0,065

Fonte: elaboração própria.

O elevado **grau de excentricidade** das elipses formadas pelas categorias dos menos e dos mais escolarizados, respectivamente 0,796 e 0,759 (ver Tabela 2.1), constitui um indício de que os indivíduos nessas subnuvens são relativamente homogêneos ao longo de um eixo (no caso, o eixo 1) e heterogêneos em relação ao outro (eixo 2). Isso significa que, entre os mais escolarizados, por exemplo, há relativa homogeneidade no que se refere à elevada participação cultural e "controle" de um repertório de gostos legítimos e maior heterogeneidade no que se refere à rejeição de gostos "onívoros". Diferentemente, a elipse que inclui a subnuvem dos indivíduos com ensino médio completo (84,8% dos pontos) possui menor grau de excentricidade (0,497), assemelhando-se mais a uma esfera, o que significa que tal subnuvem é caracterizada por maior heterogeneidade, se comparada com as demais, ao longo dos dois eixos do plano fatorial em questão.

Figura 2.4 | Elipses de concentração e subnuvens resultantes da partição da nuvem global de indivíduos conforme as categorias da variável "classe social"



Fonte: elaboração própria.

Observando as elipses de concentração de quatro subnuvens formadas pela partição da nuvem global conforme as categorias "profissionais", "quadros", "tmq/stm" e "tmnq" da variável "classe", nota-se um padrão de distribuição similar àquele da variável "escolaridade": como a variável está correlacionada com os dois eixos do plano fatorial, as elipses vão se "esparramando" a partir do quadrante superior à direita até o quadrante inferior à esquerda. As elipses que englobam os pontos

das subnuvens dos profissionais, dos supervisores e de trabalhadores manuais qualificados têm formato e áreas similares, embora a primeira tenha maior grau de excentricidade do que a segunda, o que indica que a subnuvem dos profissionais é mais heterogênea em relação ao segundo eixo e mais homogênea em relação ao primeiro (de forma similar à elipse de concentração da subnuvem formada pelos trabalhadores manuais não qualificados). Por sua vez, a elipse que concentra os pontos da subnuvem dos quadros apresenta uma disposição bem diferente, com a maior excentricidade das quatro, o que indica que tal subnuvem é heterogênea em relação ao primeiro eixo e mais homogênea em relação ao segundo. Ou seja, os quadros diferem bastante em relação ao engajamento/desengajamento cultural e em relação à adesão à cultura "legítima" (eixo 1) e menos em relação ao "onivorismo" (eixo 2). É provável que tais características tenham a ver com os padrões de recrutamento dos quadros gerenciais, que apresentam maior variação em termos do perfil de renda e de escolaridade que profissionais ou trabalhadores manuais.

Além disso, a grande sobreposição entre as elipses de quadros e profissionais também coloca em dúvida uma parte da interpretação dos fatores estruturantes do segundo eixo feita acima, a que faz referência a uma oposição entre quadros e profissionais em função das distâncias entre as coordenadas de suas categorias nesse eixo.[5]

Para lidarmos com essa questão, podemos recorrer ao **teste de tipicalidade**, aplicado ao plano fatorial. Esse teste permite examinar

> a tipicalidade de uma subnuvem de *n* indivíduos com respeito à nuvem global de *N* indivíduos. O conjunto dos N indivíduos é tomado como **a população de referência**, e um subconjunto de n elementos da população de referência como a amostra... A cada amostra está associada uma subnuvem possível, portanto, $\left(\frac{N}{n}\right)$ subnuvens possíveis (Le Roux; Rouanet, 2010, p. 82, ênfases no original).

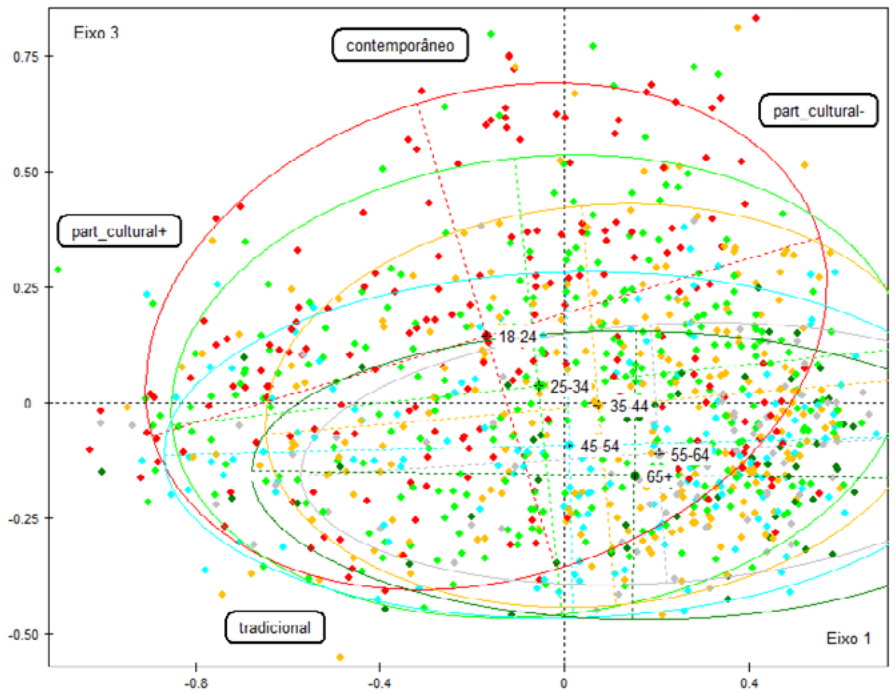
[5] Para passarmos da nuvem de indivíduos para a de categorias, basta aplicarmos uma das fórmulas de transição, que consiste na divisão da coordenada do ponto médio da subnuvem pela raiz quadrada do autovalor: para os profissionais e quadros, temos as seguintes coordenadas no primeiro e no segundo eixos, respectivamente: -1,07 e -0,74; -0,72 e 0,28.

Primeiramente, calcula-se o desvio do **ponto médio** da subnuvem em relação ao baricentro da nuvem global por meio da seguinte fórmula: $d = \sqrt{\left(\frac{\bar{y}_{l1}^k}{\sqrt{\lambda_1}}\right)^2 + \left(\frac{\bar{y}_{l2}^k}{\sqrt{\lambda_2}}\right)^2}$ ou $\sqrt{(y_{l1}^k)^2 + (y_{l2}^k)^2}$, sendo $\bar{y}_{l1}^k$ a coordenada do ponto médio da subnuvem no eixo 1 (nuvem de indivíduos); λ_1, o autovalor do eixo 1; e y_{l1}^k, a coordenada da categoria no eixo 1 (nuvem de modalidades); $\bar{y}_{l2}^k$, a coordenada do ponto médio da subnuvem no eixo 2 (nuvem de indivíduos), e assim por diante. Para a categoria "quadros", o valor da fórmula é 0,77.

Então, calculamos o teste de tipicalidade usando a seguinte fórmula: $X^2 = n \frac{N-1}{N-n} d^2$. Ou seja, $42 \frac{995-1}{995-42} (0{,}77)^2 = 25{,}9$. Com dois graus de liberdade para o valor do qui-quadrado, o resultado é estatisticamente significativo no nível de α = 0,01 (valor crítico igual a 9,21).[6] O resultado sugere que a subnuvem dos quadros é **atípica** em relação à nuvem global no plano fatorial formado pelos eixos 1 e 2, estando mais à esquerda do primeiro eixo e acima do segundo (ou seja, no quadrante da "participação cultural elevada" e "onivorismo").

[6] Para facilitar a exposição, são utilizadas as frequências absolutas não ponderadas. Utilizando os valores ponderados, o valor da fórmula é 25,01, o que não altera a interpretação.

Figura 2.5 | Elipses de concentração e subnuvens resultantes da partição da nuvem global de indivíduos conforme as categorias da variável "idade" (eixos 1 e 3)



Fonte: elaboração própria.

Considerando o plano fatorial formado pelos eixos 1 e 3, observemos as elipses de concentração das subnuvens resultantes da partição da nuvem global a partir das categorias da variável "idade". Lembremos que a inspeção visual das localizações dos pontos das categorias dessa variável e o cálculo das distâncias entre suas coordenadas na nuvem de modalidades indicaram a existência de uma correlação entre o terceiro eixo e essa variável. Observando a Figura 2.5, notamos que as elipses de concentração se distribuem do quadrante inferior à direita ("baixa participação cultural"/"gosto clássico"), onde estão as subnuvens dos indivíduos mais idosos, até o quadrante superior à esquerda ("elevada participação cultural"/"gosto contemporâneo"), onde estão as subnuvens

dos indivíduos mais jovens. Esse padrão de distribuição indica, como sabemos, que a variável está correlacionada com os dois eixos que formam o plano fatorial. A elipse que inclui os pontos da categoria "18-24" tem excentricidade de 0,716 e, considerando que ela se estende por quase todo o eixo 1, percebe-se que existe maior heterogeneidade nessa subnuvem em relação a esse eixo (participação cultural) do que em relação ao terceiro. O mesmo ocorre, de modo ainda mais acentuado, em relação à elipse de concentração dos pontos da categoria "65+". Com excentricidade de 0,926, tem um formato bastante ovalado, estendendo-se pelo eixo 1. Isso significa que há elevada heterogeneidade em relação ao eixo 1, mas relativamente pouca em relação ao terceiro (argumentos similares se aplicam às elipses de concentração das categorias "45-54" e "55-64"). Por sua vez, as elipses das subnuvens das categorias "25-34" e "35-44" ocupam porções intermediárias do plano fatorial, sendo que a primeira tem uma maior inclinação em direção ao quadrante superior à esquerda. O resultado do **teste de tipicalidade**, considerando o plano fatorial, da subnuvem que inclui os indivíduos com idade entre 25 e 34 é estaticamente significativo (igual a 16,9) para o valor crítico do qui quadrado ao nível de significância de 0,01, o que significa que a posição dessa subnuvem no plano fatorial é atípica em relação à nuvem global. Em outras palavras, a referida subnuvem diferencia-se em termos das duas dimensões do plano fatorial, estando mais para o quadrante "participação cultural+"/"gosto contemporâneo".

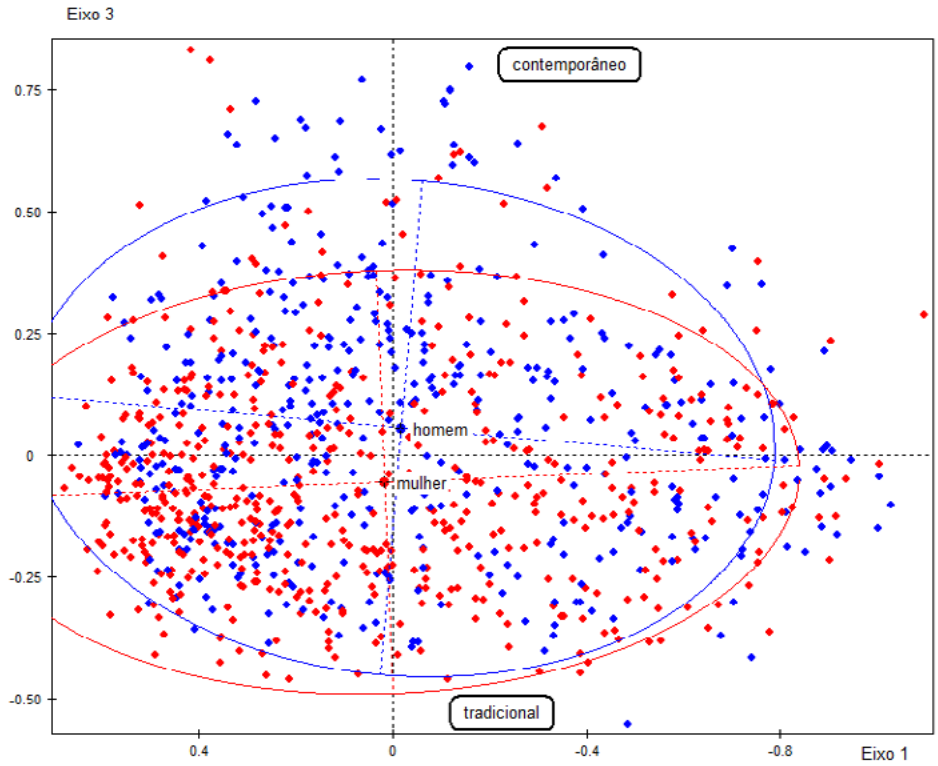
Retomando, neste ponto, a questão que fizemos anteriormente sobre a associação entre a variável sexo e os contrastes no terceiro eixo, podemos questionar se a distância entre as coordenadas de duas categorias suplementares na nuvem de modalidades pode ser considerada significativa, ainda que seja menor que o valor mínimo convencionalmente estabelecido (≥ 0,5). Vimos que esse é o caso das categorias da variável "sexo", cuja distância é de 0,46 (um pouco abaixo do considerado significativo). As elipses de concentração das categorias "homem" e "mulher" se sobrepõem quase que completamente no plano fatorial formado pelos eixos 1 e 2, o que evidencia que a variável sexo não está correlacionada com nenhum dos eixos que o constituem. Em outras palavras, saber que um indivíduo é homem ou mulher não nos

ajuda a prever sua posição em qualquer dos quadrantes desse plano. Diferentemente, no plano fatorial formado pelos eixos 1 e 3 (Figura 2.6), ocorre também uma significativa sobreposição, mas a elipse de concentração da subnuvem formada pelos homens avança relativamente mais em relação à parte superior do eixo 3 ("cultura contemporânea"), enquanto a das mulheres, por sua vez, ocupa uma porção maior da região inferior do referido eixo ("cultura tradicional").

Podemos fazer uso novamente do teste de tipicalidade, mas, desta vez, **aplicado ao eixo principal apenas** (no caso, o eixo 3). A fórmula para o teste de tipicalidade para um eixo principal é a seguinte: $Z = \sqrt{n \frac{N-1}{N-n}} \frac{\bar{y}_l^k}{\sqrt{\lambda}}$. Para a categoria "homem", esse valor é $\sqrt{447 \frac{995-1}{995-447}} \frac{0{,}056}{\sqrt{0{,}0596}} = 6{,}5$. Este valor é superior a **2,576**, valor crítico de *z* para um nível de significância para o teste bilateral (α) igual a 0,01 (ou 0,005 para o teste unilateral). Isso significa que a subnuvem dos homens é **atípica** no eixo 3, estando mais para o lado "contemporâneo" (acima do eixo).[7] Notemos que, **quanto maior o tamanho da amostra, mais provável será obter resultados estatisticamente significativos para esse teste**. Por isso, é fundamental levar em conta outros parâmetros para o exame dos fatores estruturantes do espaço fatorial.

[7] Os resultados são igualmente significativos para a categoria das mulheres (Z = 8,03), pois a localização do ponto médio da subnuvem é idêntico, mas do lado oposto do eixo, e há um número maior de mulheres. O resultado do teste, nesse caso, deve ser interpretado sublinhando que a subnuvem das mulheres é igualmente atípica, na porção inferior do eixo 3. Os cálculos foram efetuados considerando as frequências absolutas não ponderadas, de modo a facilitar a exposição. Os resultados não diferem significativamente dos apresentados quando são considerados os valores ponderados.

Figura 2.6 | Elipses de concentração e subnuvens resultantes da partição da nuvem global de indivíduos conforme as categorias da variável "sexo" (eixos 1 e 3)



Fonte: elaboração própria.

As elipses de concentração nos permitem, portanto, visualizar a dispersão das subnuvens de pontos (ao longo de um eixo e em um plano fatorial), resultantes da partição da nuvem global conforme as categorias de uma variável suplementar. A observação das elipses e a análise de suas características (formato, tamanho, excentricidade) são passos adicionais importantes em relação ao cálculo das distâncias entre as coordenadas fatoriais das categorias suplementares na nuvem de modalidades.

Por fim, podemos decompor a variância das subnuvens resultantes da partição da nuvem de indivíduos conforme as categorias da variável

suplementar: a variância **entre** as subnuvens (ou internuvens) e a variância **dentro** delas (ou intranuvens). Para entendermos esses conceitos, retomemos a definição de variância da nuvem, entendida como a soma das distâncias quadráticas dos pontos da nuvem em relação ao ponto médio dividida pelo número total de pontos, sendo o ponto médio de uma nuvem calculado a partir da seguinte fórmula: $G = \sum \frac{M^i}{n}$, em que M^i é um ponto qualquer e n, a quantidade de pontos. Ou seja, G é "a média dos pontos de uma nuvem" (Le Roux; Rouanet, 2010, p. 18). Se particionarmos uma nuvem conforme as categorias de uma variável suplementar (por exemplo, com três modalidades), teremos três subnuvens. Uma subnuvem é como uma nuvem qualquer, "com um *peso* e um ponto médio".

Os pontos médios das subnuvens constituem uma "nova nuvem", denominada "internuvem" [*between-cloud*], que é, conforme a definição dos autores mencionados, uma "**nuvem ponderada**, ou seja, cada um de seus pontos é considerado com o peso da subnuvem de onde se originam". Por consequência, a variância da **internuvem** é a "**média ponderada** das distâncias quadráticas de seus pontos ao ponto médio" da nuvem global". Trata-se aqui de uma medida de quanto os pontos médios das subnuvens se afastam em relação ao ponto médio da nuvem global, quer dizer, uma medida dos desvios entre os pontos das diferentes subnuvens ao longo de um eixo ou de um plano fatorial. Quando maior a variância internuvem, mais os pontos médios das subnuvens estarão distantes do ponto médio da nuvem global. Por sua vez, "a **variância intranuvem** [*within-cloud*] associada com uma partição será definida como a média ponderada das variâncias das subnuvens da partição." (Le Roux; Rouanet, 2010, p. 19-21, ênfase no original). A variância intranuvem, portanto, é uma medida da **homogeneidade** de uma subnuvem ao longo de um eixo ou de um plano fatorial. Valores menores da variância intranuvem indicam que seus pontos estão relativamente próximos entre si no eixo ou no plano fatorial e também próximos ao ponto médio da subnuvem, o que significa uma homogeneidade elevada dos indivíduos no que se refere aos contrastes ou oposições do eixo ou do plano fatorial.

A razão entre a variância internuvem e a variância total (calculada a partir da soma dos valores das variâncias inter e intranuvem) é denominada por η^2, *eta* quadrado, medida análoga ao r^2 nas análises de correlação e de regressão (Hjellbrekke, 2019, p. 75). Ou seja, quanto maior o η^2, mais fortemente uma variável estará correlacionada com um determinado eixo.

Para obter a **dupla decomposição da variância da nuvem global**, sigo aqui as indicações contidas em Hjellbrekke (2019, p. 74): as localizações dos indivíduos em cada eixo são tomadas como novas variáveis em uma ANOVA unidirecional; a soma total dos quadrados dividida pelo total de graus de liberdade é igual à inércia do eixo (no caso, a variável endógena) desde que sejam incluídos todos os casos ativos. A Tabela 2.2, a seguir, apresenta os valores das variâncias internuvem ou intergrupo (a soma dos quadrados da regressão) e intranuvem ou intragrupo (soma dos quadrados dos resíduos), o valor total (resultante da soma dos dois tipos de variância) e o valor do *eta* quadrado, que pode ser interpretado como a porção da variância da variável dependente que pode ser "explicada" por uma variável exógena ou preditora.

Tabela 2.2 | Valores da variância internuvem, da variância intranuvem e do eta quadrado conforme as categorias de diferentes variáveis nos três primeiros eixos

Variável		Eixo 1	Eixo 2	Eixo 3
Idade				
	Internuvem	12,94*	2,74	8,52*
	Intranuvem	152,76	108,28	50,75
	Total	165,70	111,03	59,27
	η^2	0,08	0,02	0,14
Escolaridade				
	Internuvem	45,82*	4,87*	1,35
	Intranuvem	119,88	106,16	57,92
	Total	165,70	111,03	59,27
	η^2	0,28	0,04	0,02

Variável		Eixo 1	Eixo 2	Eixo 3
Classe social				
	Internuvem	32,15*	6,31*	1,65
	Intranuvem	133,55	104,72	57,62
	Total	165,70	111,03	59,27
	η^2	0,19	0,06	0,03
Origem social				
	Internuvem	13,68*	4,50*	1,59
	Intranuvem	152,02	106,53	57,68
	Total	165,70	111,03	59,27
	η^2	0,08	0,04	0,03
Sexo				
	Internuvem	0,30	0,00	3,09*
	Intranuvem	165,4	111,03	56,18
	Total	165,70	111,03	59,27
	η^2	0,00	0,00	0,05
Origem educacional				
	Internuvem	32,14*	2,42	0,94
	Intranuvem	133,56	108,61	58,33
	Total	165,70	111,03	59,27
	η^2	0,19	0,02	0,02
Renda total *per capita*				
	Internuvem	29,82*	1,53	1,05
	Intranuvem	135,88	109,50	58,22
	Total	165,70	111,03	59,27
	η^2	0,18	0,01	0,02

Fonte: elaboração própria.
* significativo ao nível .000 (valor-p).

Notemos, inicialmente, que a variância intranuvem é, para todas as variáveis, mais elevada do que a variância internuvem, o que é um resultado esperado. A variável escolaridade é o principal fator estruturante no primeiro eixo, "explicando" 28% da variância (no caso, as variações nas posições dos indivíduos ao longo do eixo), seguida pelas variáveis classe social (agregado ocupacional do indivíduo) e origem educacional (maior escolaridade do pai e/ou da mãe), com 19% da variância, mais até do que a variável renda (rendimentos totais no domicílio em salários-mínimos *per capita*). A variável origem social é também um fator estruturante no primeiro eixo, embora "explicando" uma proporção menor da variância (8%). Algumas dessas variáveis são também fatores estruturantes no segundo eixo, sendo classe social a principal delas (6% da variância).

Em relação à variável sexo, **os resultados da dupla decomposição da variância confirmam nossas análises prévias**: a variável não tem qualquer peso na "explicação" da variância nos dois primeiros eixos (sendo praticamente nulo) e possui o segundo maior valor (ainda que pequeno) no terceiro eixo (5%).

Idade, por sua vez, é a variável que "explica" a maior proporção da variância do terceiro eixo (14%), entre as variáveis consideradas. Voltando à Figura 2.5, vemos como as elipses de concentração dessa variável se distribuem ao longo do eixo segundo um padrão claramente discernível: os mais idosos, na parte inferior; os mais jovens, na parte superior, indicando maior ou menor aderência à cultura tradicional ou à cultura contemporânea. A dupla decomposição da variância dessa variável nos três eixos nos dá mais indícios de que, de fato, conhecendo a idade de um indivíduo, temos uma boa probabilidade de inferirmos sua localização no eixo, ou seja, suas preferências e práticas culturais; ou, ao revés, que sabendo a localização de um indivíduo no eixo, podemos inferir, com boa probabilidade de acerto, sua idade de forma aproximada (ou, pelo menos, sua faixa etária).

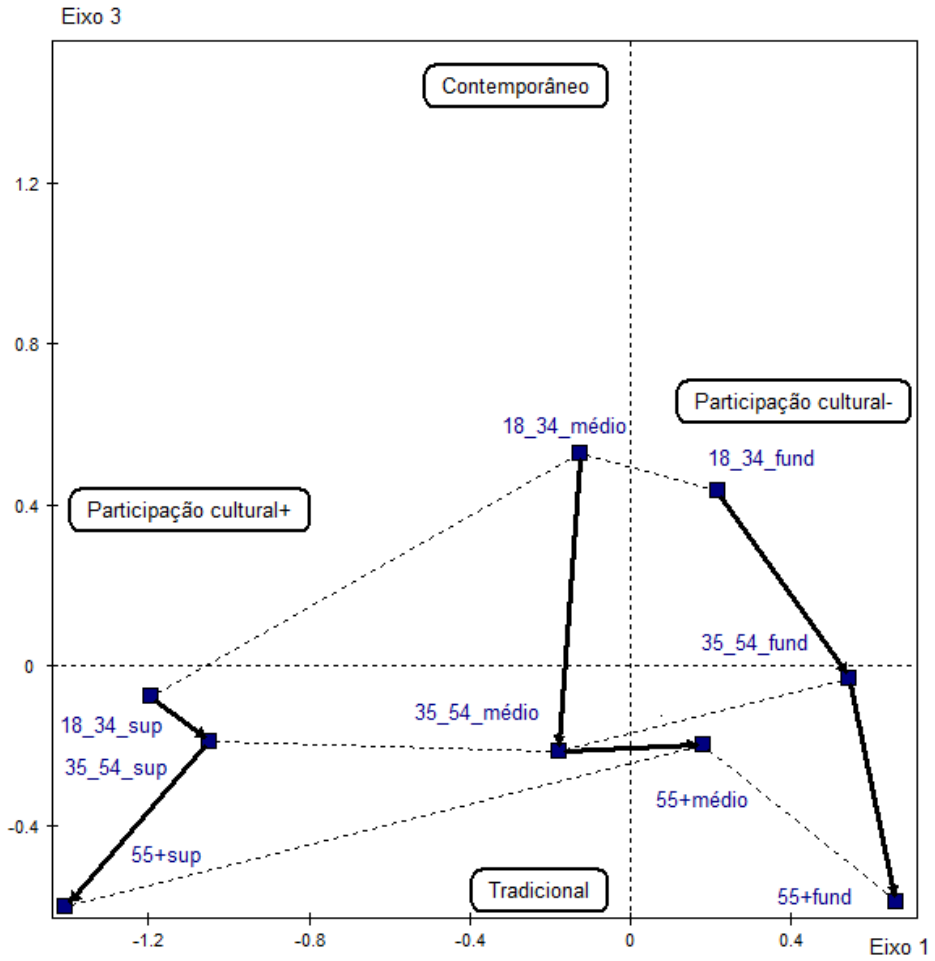
O cruzamento de dois fatores estruturantes

A partir da análise anterior, sabemos que a idade e a escolaridade, entre outros, são fatores estruturantes das afinidades e contrastes entre as práticas culturais dos agentes. Em outras palavras, conhecendo a idade ou a escolaridade de um indivíduo, é possível situá-lo no espaço das práticas culturais ou, então, sabendo sua localização no referido espaço, inferir algumas de suas características. Outra questão importante é saber se tais fatores interagem entre si para produzir efeitos específicos sobre as práticas. Por exemplo, sabemos que o terceiro eixo opõe práticas e gostos contemporâneos, de um lado, e práticas e gostos tradicionais, de outro, e que tal eixo é fortemente estruturado pela idade dos agentes, estando os mais jovens mais próximos à cultura contemporânea, e os mais idosos, à cultura tradicional. Será que tal contraste apreende adequadamente a relação dos indivíduos com os bens culturais quando consideramos simultaneamente idade e escolaridade? Ou melhor: há diferenças entre jovens e idosos, no que se refere à relação com a cultura tradicional ou contemporânea, conforme seu nível de capital escolar?

Como argumenta Le Roux (2014), quando fazemos o cruzamento de dois fatores estruturantes A x B, é importante diferenciar os **efeitos principais** (os efeitos de A e os efeitos de B) dos **efeitos de interação**. Os efeitos principais podem ser observados quando são projetadas as modalidades de variáveis suplementares ao longo de um eixo (Figura 2.7). Na Figura 2.7, os efeitos principais são indicados pelas setas conectando as modalidades das variáveis escolaridade e idade: as modalidades da primeira variável se distribuem ao longo do eixo horizontal e aquelas da segunda, ao longo do eixo vertical.[8]

[8] Para facilitar a exposição dos argumentos, as variáveis foram recodificadas de maneira a reduzir o número de modalidades. A variável escolaridade possui três modalidades: "fund", que inclui os indivíduos com até o ensino fundamental completo; "médio", que inclui os indivíduos com ensino médio completo; e "superior", que inclui aqueles com ensino superior completo ou incompleto. A variável idade possui também três categorias: indivíduos com idade entre 18 e 34 anos; entre 35 e 54 anos; e, por fim, com 55 anos ou mais.

Figura 2.7 | Efeitos principais e de interação das variáveis "escolaridade" e "idade" no plano fatorial formado pelos eixos 1 e 3



Fonte: elaboração própria.

Uma nuvem construída a partir do cruzamento de dois fatores não apresenta interação "se os efeitos de A dentro de B forem os mesmos para os diferentes níveis de B ou, de forma equivalente, se os efeitos de B dentro de A forem os mesmos para os diferentes níveis de A" (Le Roux, 2014, p. 197). Consequentemente, a nuvem formada pelos pontos resultantes

do cruzamento desses fatores teria o formato de um paralelogramo na ausência de interação.

Para examinar os possíveis efeitos de interação, projetamos agora as nove modalidades resultantes do cruzamento das duas variáveis (Figura 2.7). Os efeitos de idade dentro de escolaridade estão indicados por setas; já os efeitos de escolaridade dentro de idade, por linhas tracejadas. Notemos que a distância que separa as modalidades da variável idade varia fortemente conforme o nível de escolaridade: é relativamente pequena entre aqueles com ensino superior (as modalidades estão localizadas no quadrante inferior à esquerda) se comparada com a distância que separa as modalidades de idade entre os menos escolarizados. A distância no eixo vertical entre as modalidades "18_34_sup" (jovens entre 18 e 34 anos com ensino superior) e "55+sup" (indivíduos com 55 anos ou mais com ensino superior) é igual a 0,52 [-0,6 - (-0,08)]; entre "18_34_fund" (indivíduos entre 18 e 34 anos com até o ensino fundamental completo) e "55+ fund" (indivíduos com 55 anos ou mais com a mesma escolaridade) é igual a 1,02 [0,43 - (-0,59)]. A distância é também relativamente grande entre os indivíduos com escolaridade média e idades entre 18 e 34 anos, de um lado, e 35 a 54 anos, de outro, sendo igual a 0, 74 [(053 – (-0,21)]. Como as distâncias entre as modalidades da variável idade não são as mesmas para as diferentes categorias de escolaridade, é possível afirmar que variam os efeitos de uma variável nos diferentes níveis da outra, havendo, portanto, algum efeito interativo entre elas. Outra maneira de interpretar esses resultados consiste em afirmar que as diferenças entre mais jovens e idosos, no que se refere aos contrastes do eixo vertical, são menores entre os mais escolarizados, e maiores entre os menos.

Interpretações similares podem ser feitas em relação aos efeitos da variável escolaridade nos diferentes níveis da variável idade. Notemos, por exemplo, que a distância entre os mais jovens (entre 18 e 34 anos) nos três níveis de escolaridade é menor do que aquela entre os mais idosos (55 anos ou mais). A distância que separa as categorias "18_34_fund" e "18_34_sup" é igual a 1,41 [0,22 – (-1,19)]; já aquela que separa as categorias "55+sup" e "55+fund" é igual a 2,07 [0,66 – (-1,41)]. Ou seja, há menos diferenças entre os jovens do que entre os mais idosos, independentemente do nível de capital escolar, no que se refere à participação cultural.

Conclusões

Uma das propriedades mais importantes da ACM consiste em que as nuvens de modalidades e de indivíduos possuem a mesma dimensionalidade, sendo possível passar de uma à outra por meio de algumas fórmulas de transição. É possível, nesse sentido, conhecer as coordenadas de um indivíduo no plano fatorial a partir de seu padrão de respostas, assim como é possível saber o ponto médio de uma subnuvem (conjunto de indivíduos que "escolheram" determinada categoria) a partir da coordenada fatorial dessa categoria na nuvem de modalidades. Tal propriedade permite, também, que as chamadas modalidades suplementares (que não participam da construção dos eixos e da definição das distâncias entre indivíduos e categorias ativas) sejam projetadas nos planos fatoriais construídos a partir dos eixos retidos para interpretação. Ou seja, uma vez construídas as nuvens, as categorias suplementares podem ser projetadas na nuvem de categorias, assim como os indivíduos suplementares podem ser projetados na nuvem de indivíduos.

A inclusão das variáveis suplementares constitui uma técnica da AGD para lidar com os chamados **fatores estruturantes**, ou seja, as "variáveis relevantes que descrevem os dois conjuntos básicos [das linhas e das colunas] e que não servem para construir o espaço geométrico" (Le Roux; Rouanet, 2004, p. 251). Diferentemente de outras técnicas multivariadas, a ACM não pressupõe a diferenciação das variáveis entre dependentes e independentes. No caso da ACM, se temos interesse em saber se idade ou escolaridade estão associadas com práticas culturais, então é melhor inserir tais variáveis como suplementares uma vez construídos os "mapas culturais".

A projeção das categorias suplementares na nuvem de modalidades nos permite enriquecer a interpretação dos eixos. Num primeiro momento, **a inspeção visual das localizações das categorias suplementares**, acompanhada do **cálculo das distâncias entre suas coordenadas ao longo dos eixos** que formam o plano fatorial, nos dão indícios das associações entre as modalidades e os eixos.

É necessário complementar essa análise inicial com outras duas: a observação das elipses de concentração que incluem as subnuvens resultantes da partição da nuvem global conforme as categorias das variáveis suplementares, atentando para o formato e localização dessas elipses nos quadrantes do espaço fatorial, além dos níveis de sobreposição entre elas; a dupla decomposição da variância das subnuvens (variância **internuvem** e **intranuvem**).

Por fim, é importante examinar se os efeitos principais dos fatores estruturantes variam quando comparamos diferentes subgrupos de indivíduos. Assim, é possível investigar, por exemplo, se as práticas e gostos culturais de homens e mulheres ou de jovens e idosos apresentam diferenças importantes conforme o nível de capital escolar.

Capítulo 3

Analisando subgrupos: a análise de classes específicas como uma variante da ACM

No capítulo anterior, vimos que a análise dos dados estruturados tem uma orientação preditiva: sabendo a localização de um indivíduo no plano fatorial, temos a possibilidade de prever o valor que possui em algum fator estruturante (exemplo: idade) ou, então, sabendo tal valor, inferir sua provável localização no espaço cartesiano. Um indivíduo jovem, profissional, altamente escolarizado, por exemplo, tem grande probabilidade de ocupar o quadrante do plano fatorial (formado pelos eixos 1 e 2) caracterizado por elevada participação cultural, por gostos "legítimos" e pela evitação do "onivorismo". Sabemos, portanto,

que as variáveis idade, classe social, origem social, entre outras, estão associadas a um ou mais dos três eixos.

A partir daí, poderíamos perguntar se as oposições que caracterizam a nuvem global se reproduzem quando consideramos grupos específicos? O contraste entre maior e menor participação cultural caracteriza igualmente os mais idosos e os mais jovens? E o que dizer do contraste entre "gosto tradicional" e "gosto contemporâneo": estrutura igualmente os "subespaços" dos mais escolarizados e dos menos escolarizados?

Tais questões podem ser respondidas recorrendo à **análise de classes específicas**, doravante ACE: trata-se de uma variante da ACM que possibilita examinar "uma classe (subconjunto) dos indivíduos com referência ao conjunto total de indivíduos (ativos), ou seja, para determinar as características específicas dessa classe" (Le Roux; Rouanet, 2010, p. 64). A ACE possibilita estudar, portanto, uma subnuvem de indivíduos, "embora ainda preservando as distâncias definidas para a nuvem inteira". Este é um ponto importante: **a ACE difere da ACM convencional aplicada a um subgrupo de indivíduos**, "em que as distâncias entre os pontos e os pesos dos pontos são definidos a partir das margens da subtabela de dados". Diferentemente, na ACE, "as distâncias e os pesos são definidos a partir da tabela inicial completa" (Lebaron; Bonnet, 2019, p. 375).

A ACE é um recurso metodológico bastante útil à tarefa de investigar

> espaços sociais 'entranhados' [*embedded*]: com a condição de que a referência completa tenha sido inicialmente construída, todos os outros espaços podem ser analisados como subespaços, permanecendo, ainda, situados **dentro** desse espaço global, e estudados enquanto tais (Lebaron; Bonnet, 2019, p. 359, ênfase no original).

Dessa forma, podemos responder a questões do tipo: "em que medida as estruturas internas a um subgrupo são similares às estruturas no espaço global? A dimensionalidade na subamostra é a mesma que na amostra

completa? Os eixos e seu ordenamento são os mesmos? Se não, quais são as implicações analíticas?" (Hjellbrekke, 2019, p. 101).

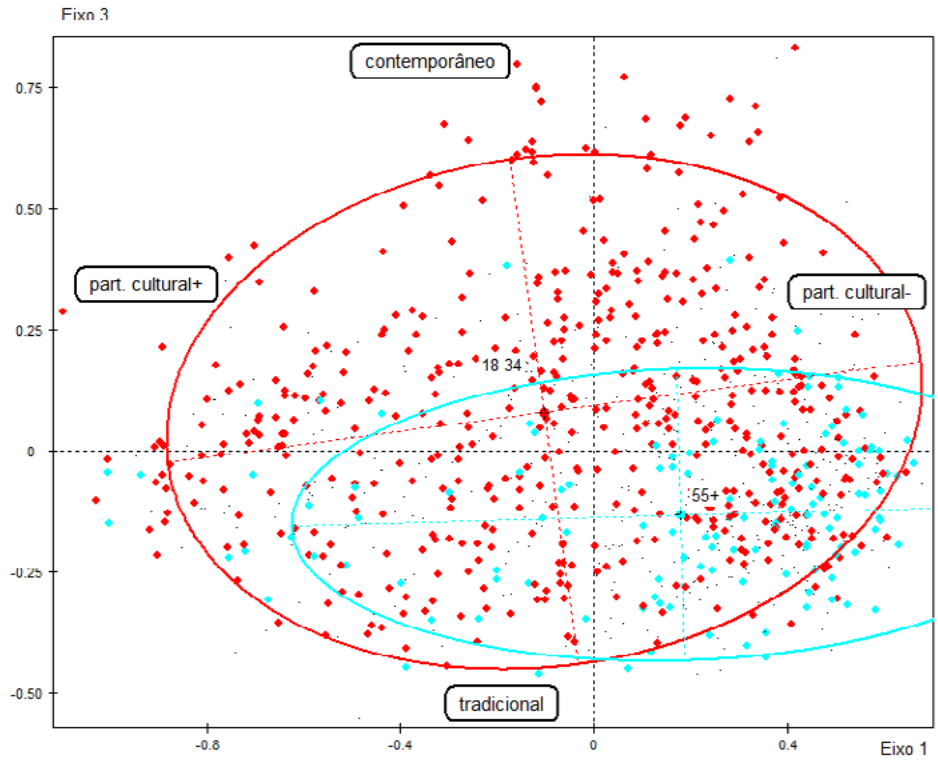
Anteriormente, vimos que o espaço global das práticas e dos gostos culturais é caracterizado por três contrastes principais: entre engajamento *versus* desengajamento cultural, no primeiro eixo; entre um padrão de gostos marcado pelo onivorismo *versus* um padrão de gostos mais seletivo, no segundo eixo; e, por fim, um contraste entre gostos tradicionais *versus* gostos contemporâneos. Será que tais oposições estruturam internamente o subespaço formado pelas pessoas mais jovens?[9] E aquele formado pelos mais velhos?

Para exemplificar o uso da ACE, será feita uma comparação entre duas subnuvens de indivíduos, construídas a partir da partição da nuvem global conforme duas categorias da variável "idade", a dos indivíduos entre 18 e 24 anos (189 casos) e aquela dos indivíduos com mais de 55 anos (128 casos).[10] As elipses de concentração das referidas subnuvens podem ser observadas na Figura 3.1, a seguir. Notemos que os pontos médios das categorias estão em quadrantes distintos, indicando, como já sabemos, uma diferenciação ao longo dos dois eixos que constituem o plano fatorial formado pelos eixos 1 e 3. Sabemos, portanto, que os indivíduos mais jovens estão mais concentrados no quadrante caracterizado por maior participação cultural e por gostos contemporâneos, enquanto os mais velhos, no quadrante caracterizado por menor participação cultural e por gostos clássicos. Nosso objetivo agora será distinto: saber se as oposições internas a cada um desses subgrupos reproduzem as oposições no espaço global ou delas diferem?

[9] Como vários estudos evidenciam, a idade é um fator bastante correlacionado com as práticas culturais. Entre outros, Savage (2011).

[10] Uma ACE equivale a uma ACP aplicada a um subgrupo da amostra utilizando os eixos da ACM global como novas variáveis e as coordenadas dos indivíduos como valores ativos.

Figura 3.1 | Projeção das elipses de concentração e das subnuvens de duas categorias da variável "idade" no plano fatorial formado pelos eixos 1 e 3



Fonte: elaboração própria.

A seguir, é exibida a Tabela 3.1, que compara a variância total das nuvens global e específicas e os autovalores dos três primeiros eixos, com a aplicação da correção de Benzécri.

Tabela 3.1 | Variância total e específica e taxas de inércia dos três primeiros eixos nas nuvens global e específicas

	ACM Global			ACE (18-24)			ACE (55 ou mais)		
Eixo	1,182			1,271			1,035[11]		
	λ	**%**	**taxa modificada (%)**	**μ**	**%**	**taxa modificada (%)**	**μ**	**%**	**taxa modificada (%)**
1	0,1665	14,09	69,98	0,185	14,57	61,40	0,194	18,72	80,22
2	0,1116	9,44	24,91	0,125	9,80	22,68	0,100	9,62	14,41
3	0,0596	5,04	3,23	0,085	6,83	8,09	0,058	5,64	2,30

Fonte: elaboração própria.

Notemos, inicialmente, que a variância é maior no subgrupo formado pelos mais jovens e menor naquele dos mais velhos. As taxas de variância dos três primeiros eixos nos dois subespaços também diferem daquelas do espaço global: naquele formados pelos mais jovens, o terceiro eixo tem um peso maior na variância específica e, correlatamente, os dois primeiros eixos têm pesos menores (evidenciando, provavelmente, que esse subespaço tem uma estruturação mais complexa); naquele dos mais velhos, diferentemente, o primeiro eixo tem um peso maior na variância específica e os dois seguintes, pesos menores, em comparação com o que se observa na nuvem global (indicando que esse subespaço é fortemente estruturado, provavelmente, por uma única oposição apreendida no primeiro eixo)[12].

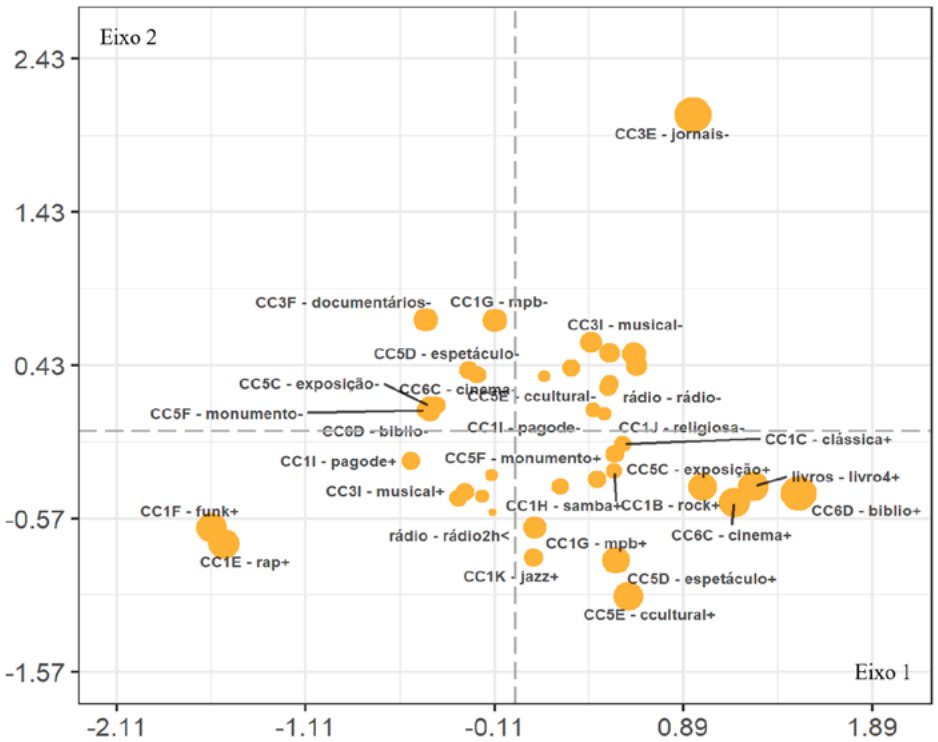
[11] A fórmula para o cálculo da variância específica é a seguinte: $V_{spe} = \frac{1}{Q}\sum_{k\in K}\frac{f_k\,(1-f_k)}{F_k}$, em que F_k é a frequência relativa da categoria k na tabela global e f_k, a frequência relativa da categoria k na subtabela.

[12] Por isso, na exposição a seguir, o subespaço constituído pelos mais idosos será examinado em termos dos contrastes ou afinidades observados nos eixos 1 e 2.

Os contrastes em cada eixo também apresentam diferenças. Para a interpretação dos eixos, consideramos, como indicado anteriormente, apenas as modalidades com contribuições acima da média. Em relação ao primeiro eixo (horizontal) da ACE dos mais jovens (Figura 3.2), há 24 modalidades com contribuições significativas, que somam 70% da variância do eixo. Tal eixo opõe, de forma similar à nuvem global, a participação cultural *versus* não participação cultural, como se depreende da observação da localização das modalidades de uso de equipamentos culturais e frequência de leitura de livros.[13] Além disso – e aí reside uma diferença importante em relação ao primeiro eixo da nuvem global – há um contraste baseado no gosto por gêneros musicais e televisivos: à esquerda do eixo, um "repertório legítimo" que inclui os gostos pelo *rock* e música clássica, além da rejeição a novelas, a programas de variedades e a telejornais; e também à música romântica, à música religiosa e ao pagode. Do outro lado do eixo, estão as modalidades que indicam os gostos pelo *rap*, *funk* e pagode, que, combinados com a menor participação na visitação a monumentos, cinemas, bibliotecas e exposições, evidenciam, em seu conjunto, um repertório de práticas mais distante da "cultura erudita".

[13] O SPAD não possui um pacote específico para a ACE. Por isso, utilizo um *script* do R que opera como uma interface no SPAD. Embora esse procedimento forneça todos os parâmetros necessários para a interpretação dos resultados, não é possível produzir as nuvens de modalidades e de indivíduos da mesma forma que na ACM padrão ou específica. No caso da ACE, temos representações gráficas dos planos fatoriais dos três primeiros eixos, geradas nas próprias planilhas de resultados, que não são editáveis. Daí as diferenças no estilo de apresentação dos planos fatoriais examinados neste capítulo. O *script* do R foi produzido por Brigitte Le Roux. Agradeço a ela pela possibilidade de utilizar o *script* nesta análise, assim como a Virgílio Borges Pereira e a Lucas Page Pereira pelos ensinamentos quanto ao modo correto de utilização do *script* no *software* SPAD.

Figura 3.2 | Projeção das categorias ativas da ACE da subnuvem dos mais jovens no plano formado pelos eixos 1 e 2



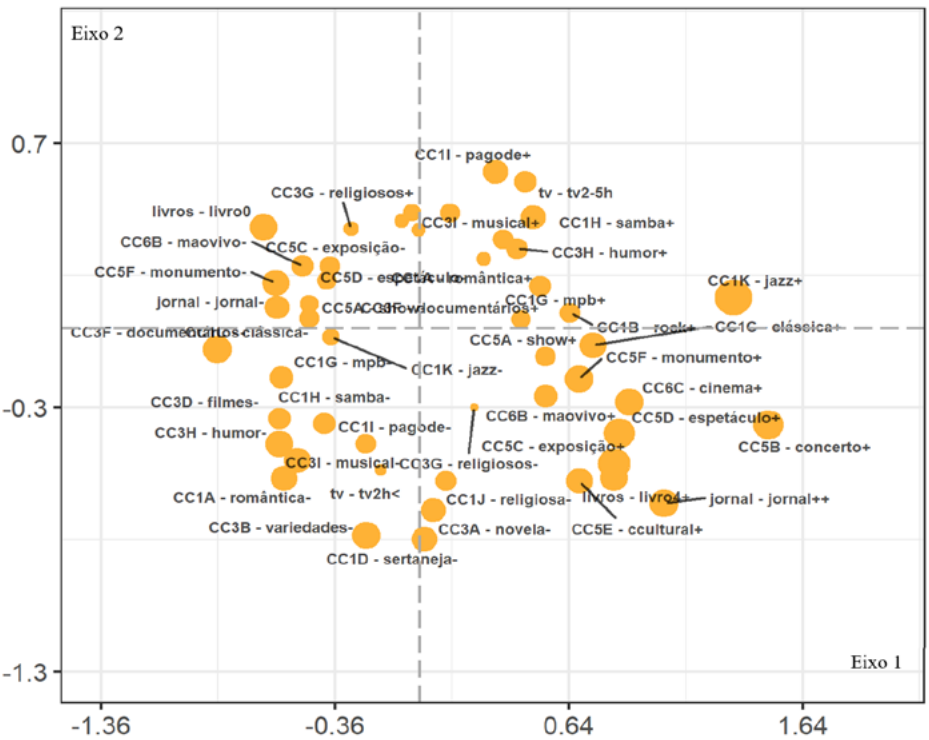
Fonte: elaboração própria.

Na segunda ACE (referente à subnuvem formada pelos mais velhos), o primeiro eixo (horizontal), de forma similar ao que ocorre no espaço global, opõe, em maior medida, as modalidades que indicam elevada participação cultural (visitação a exposições, monumentos e centros culturais; ida ao cinema, espetáculos, bar com música ao vivo, a shows e a concertos; assim como maior frequência de leitura de livros e de jornais), à direita do eixo, àquelas que indicam ausência dela, à esquerda (Figura 3.3).[14] Há também, em menor medida, um contraste entre a aderência a

[14] Há 33 modalidades com contribuições acima da média, alcançando 83% da variância total (46% se referem às variáveis de participação; 37%, às de gosto).

um repertório musical "legítimo" (caracterizado pelos gostos por *jazz*, música clássica, MPB e *rock*), além do gosto por documentários (TV) à direita, em oposição à rejeição a tal repertório, à esquerda. Vale notar, ainda, que esse eixo tem um peso relativamente grande na variância específica (80%), evidenciando aqui que a oposição entre participação *versus* não participação e aquela entre aderência *versus* rejeição ao "gosto legítimo" estrutura mais fortemente a subnuvem dos mais velhos do que aquela dos mais jovens ou a nuvem global.

Figura 3.3 | Projeção das categorias ativas da ACE da subnuvem dos mais idosos no plano fatorial formado pelos eixos 1 e 2

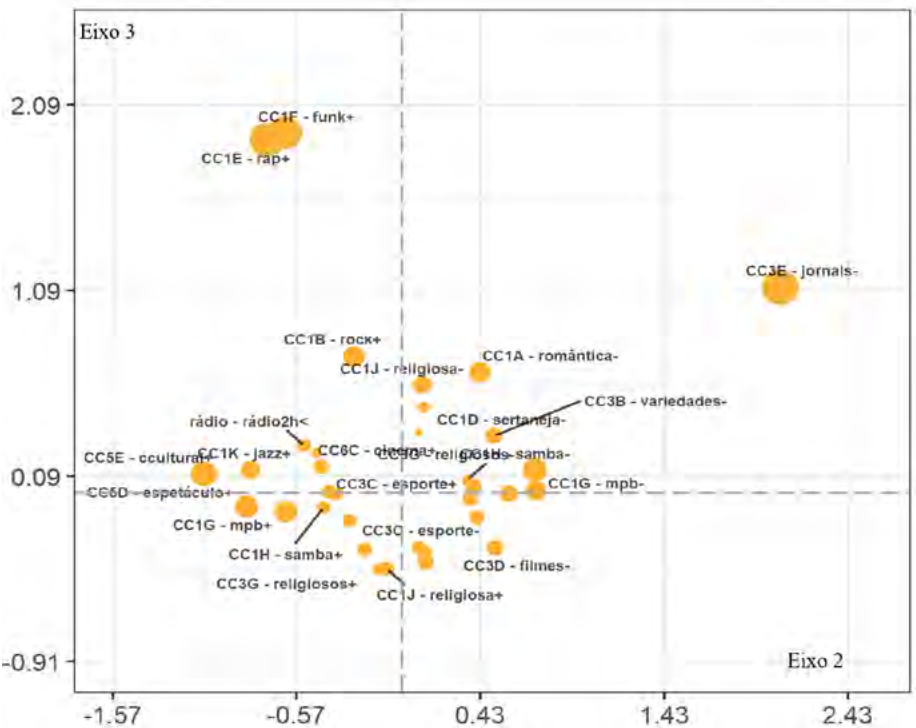


Fonte: elaboração própria.

O segundo eixo (horizontal) na subnuvem formado pelos indivíduos mais jovens (primeira ACE) revela um contraste similar àquele do espaço global, pois opõe modalidades que indicam gostos a outras que indicam

rejeições diversas. São 23 modalidades com contribuições acima da média, que somam 78% da variância do eixo (Figura 3.4). À esquerda do segundo eixo, vemos as modalidades que indicam o gosto por diferentes gêneros de televisão e de música (possivelmente evidenciando uma orientação "onívora" no consumo cultural, marcada por gostos que "atravessam" certas fronteiras culturais, como entre o *rap* e a MPB). Nessa região do eixo, estão também localizadas as modalidades de frequência a espetáculos, centros cultuais e cinemas, evidenciando – tal como a literatura sobre o tema – que um maior volume de gostos está geralmente associado à maior participação em determinadas práticas culturais. Diferentemente, à direita do eixo, estão as modalidades que indicam rejeição a diversos gêneros culturais e também menor frequência a espetáculos e centros culturais.

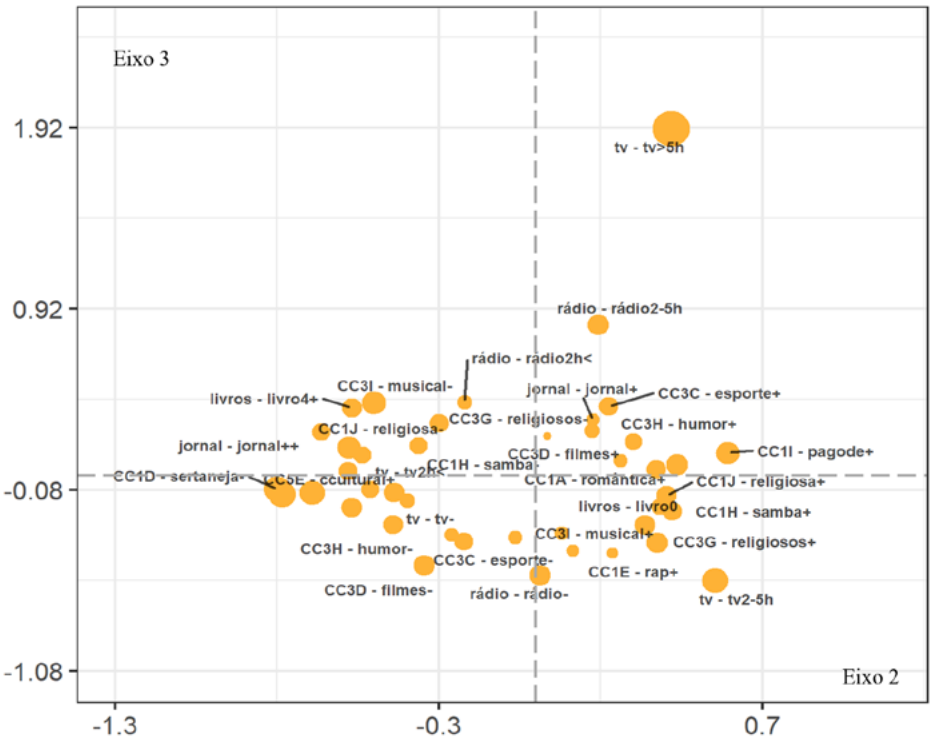
Figura 3.4 | Projeção das categorias ativas da ACE da subnuvem dos mais jovens no plano formado pelos eixos 2 e 3



Fonte: elaboração própria.

Na ACE do subgrupo dos mais idosos, o segundo eixo (horizontal) revela uma oposição um pouco diferente daquela apreendida pelo segundo eixo no espaço global (Figura 3.5). Há 28 modalidades com contribuições acima da média, somando 83% da variância do eixo. O principal contraste nesse eixo reside na oposição entre preferências e rejeições por gêneros musicais ou televisivos "populares" e, em menor medida, entre a participação *versus* não participação em práticas culturais "eruditas". À direita do eixo, estão localizadas as modalidades que indicam o gosto por música sertaneja, romântica, religiosa, pelo samba e pelo pagode; por novelas, variedades, programas musicais, religiosos e humorísticos. Nessa região do eixo, há também duas modalidades referentes à participação cultural, indicando a ausência de leitura de livros nos últimos doze meses e a prática de assistir à TV entre duas a cinco horas por dia. À esquerda do eixo, ao contrário, observamos as modalidades de rejeição a esses gêneros musicais e televisivos, além daquelas que indicam a visitação recente a exposições, espetáculos e centros culturais, a leitura de quatro livros ou mais e a prática de assistir à TV por menos de duas horas por dia.

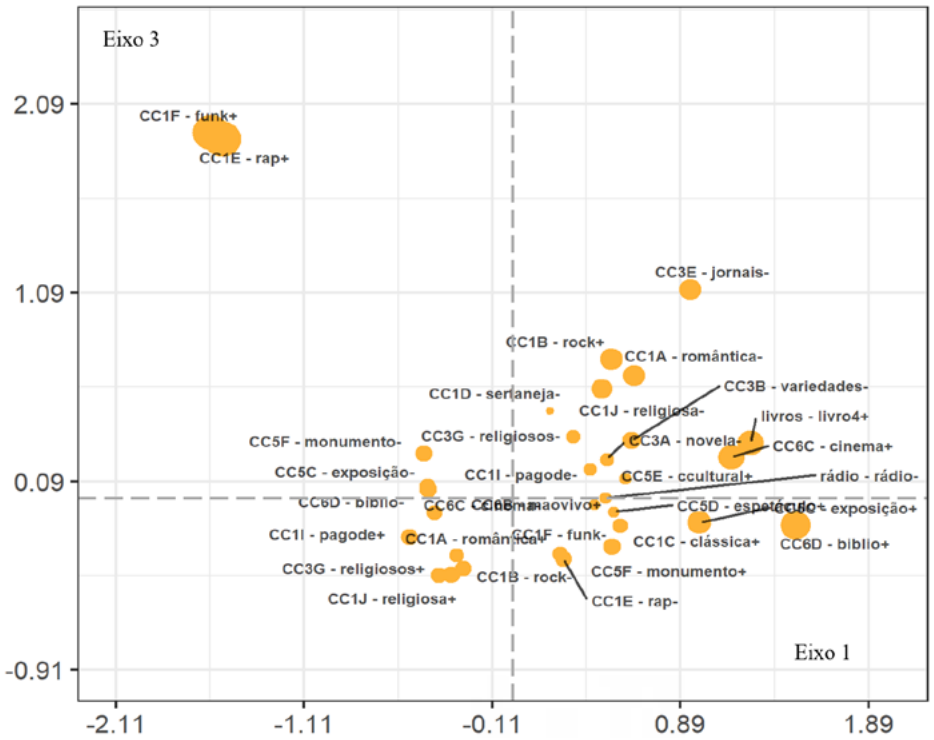
Figura 3.5 | Projeção das categorias ativas da ACE da subnuvem dos mais idosos no plano fatorial formado pelos eixos 2 e 3



Fonte: elaboração própria.

Por fim, o terceiro eixo (vertical) na ACE do subgrupo mais jovem é bastante similar àquele do espaço global ("tradicional" *versus* "contemporâneo"), embora tenha, como já vimos, maior peso na variância específica (evidenciando, portanto, que esta é uma oposição que estrutura mais fortemente essa subnuvem do que a nuvem global). Há 14 modalidades cujas contribuições somam 79% da variância do eixo (Figura 3.6). Acima da origem do plano fatorial, observamos os gostos por *rap*, *funk* e *rock*, além de rejeições à música romântica, religiosa e sertaneja, e a programas religiosos e telejornais; abaixo, notamos um padrão inverso: os gostos por programas religiosos, por música romântica e religiosa, além das rejeições ao *rap*, *funk* e *rock*.

Figura 3.6 | Projeção das categorias ativas da ACE da subnuvem dos mais jovens no plano formado pelos eixos 1 e 3



Fonte: elaboração própria.

A análise de classe específica revelou a existência de contrastes diferentes nos subgrupos formados pelos mais jovens e pelos mais velhos na comparação entre eles e com a nuvem global. A ACE forneceu evidências também de que **os subgrupos são internamente estruturados de modos diferentes**: não apenas as oposições em cada eixo são diferentes, como também difere o peso delas na variância específica. A subnuvem dos mais jovens se revelou internamente estruturada de modo mais complexo, com taxas menos desiguais de inércia entre os primeiros eixos; diferentemente, a subnuvem dos mais velhos se mostrou estruturada mais fortemente pelo primeiro eixo.

Conclusão

A ACE difere da ACM convencional aplicada a um subgrupo por preservar a estrutura do espaço global, ou seja, podemos examinar o conjunto de afinidades ou contrastes dentro de subgrupos específicos considerando suas posições relativas dentro da nuvem global. Podemos, assim, examinar as oposições internas a cada grupo em comparação com aquelas que estruturam o espaço global; comparar a dimensionalidade dos diferentes subespaços, o peso de cada eixo e a importância de diferentes modalidades para cada um deles.

Por isso, a ACE revela-se um exercício tanto mais produtivo quanto mais o exame dos subgrupos relevar dimensões, oposições ou afinidades diferentes daquelas do espaço global.

Capítulo 4

A construção de tipologias: combinando a ACM com técnicas de agrupamento

A construção de tipologias a partir da combinação de uma técnica de agrupamento (*cluster*) com a ACM tem se tornado cada vez mais comum nas ciências sociais. Tal combinação, conforme argumenta Pereira (2005, p. 198) em seu livro sobre as classes sociais e culturas de classe na cidade do Porto, tem uma função complementar: "é possível e útil, para o conhecimento [das] lógicas de estruturação e de formação [do campo das classes sociais], complementar a visão que já possuímos das suas propriedades relacionais com uma perspectiva capaz de classificar os agentes que protagonizam as referidas lógicas e, deste modo, ensaiar uma lógica e, deste

modo, ensaiar uma leitura aprofundada e tipificante dos núcleos relacionais que o constituem e formam". Ou seja, a **classificação dos casos** e a **construção de tipos**: estas são as possibilidades oferecidas pela referida combinação entre técnicas (Lebart *et al.* 1998; Le Roux; Rouanet, 2004). Pereira sustenta que

> a utilização da classificação hierárquica permite ultrapassar algumas das dificuldades decorrentes da aplicação exclusiva da ACM, nomeadamente as que passam pela interpretação de proximidades geradas por fatores para além do plano principal e pela compressão excessiva e deformação dos dados (Pereira, 2005, p. 198).

A análise feita pelo autor resultou em sete "zonas", que retratam os núcleos relacionais no espaço social portuense.[1]

Este capítulo trata da **classificação hierárquica aglomerativa** (ou ascendente), que utiliza a variância como **índice de agregação**. Esse método é também designado por **classificação euclidiana**, que se ajusta muito bem "às estruturas matemáticas da AGD". (Le Roux; Rouanet, 2004, p. 106) O resultado da classificação são "classes hierárquicas" obtidas por agrupamento e que são representadas por uma "árvore hierárquica" ou dendograma (Le Roux; Rouanet, 2004, p. 107).

Segundo Hjellbrekke (2019), a combinação entre essas técnicas nos permite responder às seguintes questões:

> i. Quão adequada é a representação nos planos fatoriais das principais estruturas nos dados?
>
> ii. Quantos grupos podemos identificar e onde se localizam nos planos fatoriais da ACM?
>
> iii. Quão fortemente a ACM permite separar os grupos assim identificados?

[1] Para análises similares, ver Bertoncelo (2016, 2019a).

Os resultados de uma análise de classificação dependem de quais técnicas de agrupamento são empregadas. A principal diferença entre as classificações hierárquicas reside no modo de construir a hierarquia: a classificação pode ser **divisiva**, que funciona de "cima para baixo", começando com os objetos a serem classificados e subdividindo-os progressivamente; ou pode ser **aglomerativa** (ascendente), que opera, diferentemente, de "baixo para cima", inicialmente considerando cada objeto como um "conjunto de classes de um elemento"; são feitas, então, sucessivas agregações até que todos os objetos sejam agrupados em uma única classe (Le Roux; Rouanet, 2004, p. 108). Ou seja, na **classificação hierárquica aglomerativa** (CHA), inicia-se com um número de classes ou agrupamentos que corresponde ao total de casos; a cada iteração (ou etapa de agrupamento), os casos vão sendo agrupados – segundo um critério qualquer – em função de suas similaridades até que todos sejam agregados numa única classe.

Quando combinada com a ACM, a CHA geralmente busca agrupar os casos de modo a minimizar a **variância intraclasse** (homogeneidade interna) e maximizar a **variância interclasse** (heterogeneidade externa). Para isso, emprega-se geralmente o método *Ward* (Hjellbrekke, 2019, p. 82). As coordenadas dos indivíduos em **todos os eixos** da ACM são usadas como *inputs* na CHA: as distâncias entre os indivíduos na nuvem euclidiana são a base para a classificação e a construção de tipos. Quando todos os eixos são incluídos na CHA, o índice de *Ward* equivale à variância da nuvem global. Por isso, essa análise é uma outra forma de particionar a variância da nuvem em termos de variâncias intraclasse e interclasse. No início, quando os casos constituem classes de um elemento (ou seja, cada caso é um *cluster*), toda a variância é de tipo **interclasse**. A cada passo da agregação, os casos (e, depois, as classes) vão sendo incluídos em agrupamentos maiores, de modo a manter tão elevada quanto possível a variância interclasse. Ao utilizar o método *Ward*, cada etapa de agregação resulta em aumento da variância intraclasse e, no fim, toda variância é desse tipo (pois há apenas um agrupamento). Essa informação é fundamental para determinar quantos agrupamentos devem ser retidos para interpretação (pois geralmente o número de classes não é definido de antemão). A observação do aumento do **eta quadrado (η^2),** calculado

a partir da divisão da variância interclasse pela variância total, ao passarmos de uma solução com **n** clusters para outra com n+1 *clusters*, é um dos fatores que devem ser considerados na tomada dessa decisão.

Idealmente, é importante termos uma solução com agrupamentos relativamente homogêneos e bem diferentes entre si, ou seja, com variância interclasse elevada e baixa variância intraclasse (portanto, um *eta* quadrado relativamente elevado). Nem sempre esse é um resultado possível de obter, sobretudo nas ciências sociais, em que geralmente temos à mão dados produzidos a partir de técnicas baseadas na observação (com pouco ou nenhum controle sobre a composição dos grupos observados), o que significa que a variância intraclasse (a heterogeneidade interna) é geralmente alta. Por isso, a questão é sempre relativa: "quando a inclusão de novos agrupamentos para de nos fornecer informações novas, relevantes?" (Hjellbrekke, 2019, p. 83). Ou: o aumento no valor do *eta* quadrado resulta em maior diferenciação entre subgrupos que é analiticamente pertinente ou os subgrupos diferenciados são muito similares entre si?

Em suma, a questão sobre quantos agrupamentos reter pode ser assim resumida:

> busca-se um ajuste entre o número de classes – preferencialmente pequeno – e a proporção da variância – preferencialmente elevada... O procedimento é encerrado quando se alcança um grau de fineza para além do qual a continuidade da subdivisão das classes não produz elevação apreciável e/ou interpretável da variância (Le Roux; Rouanet, 2004, p. 114).

Na Tabela 4.1, são apresentados os valores do *eta* quadrado, das variâncias interclasse e intraclasse e as frequências relativas dos agrupamentos resultantes de diferentes partições. Tais valores são o produto da aplicação da CHA aos resultados da ACM (convencional) feita em capítulo anterior.

Tabela 4.1 | Valores do eta ao quadrado, das variâncias interclasse e intraclasse e tamanho relativo dos agrupamentos

N (classes da partição)	η^2	Interclasse	Intraclasse	Tamanho
Partição com 2 classes	0,1044	0,1233	*Cluster* 1: 0,667	68,5%
			Cluster 2: 0,392	31,5%
Partição com 3 classes	0,1555	0,1838	*Cluster* 1: 0,365	37,7%
			Cluster 2: 0,296	34,8%
			Cluster 3: 0,336	27,5%
Partição com 4 classes	0,1833	0,2166	*Cluster* 1: 0,267	30,2%
			Cluster 2: 0,108	10,8%
			Cluster 3: 0,2709	32,2%
			Cluster 4: 0,3184	26,8%
Partição com 5 classes	**0,2006**	**0,237**	***Cluster* 1: 0,1916**	**18,6%**
			***Cluster* 2: 0,1653**	**21%**
			***Cluster* 3: 0,1130**	**10,8%**
			***Cluster* 4: 0,2404**	**30%**
			***Cluster* 5: 0,2345**	**19,6%**
Partição com 6 classes	0,2184	0,2581	*Cluster* 1: 0,1744	17%
			Cluster 2: 0,1758	22,3%
			Cluster 3: 0,1053	10,5%
			Cluster 4: 0,2405	29,6%
			Cluster 5: 0,0934	9%
			Cluster 6: 0,1343	11,6%
Partição com 10 classes	0,2681	0,3169	*	*
Partição com 50 classes	0,4259	0,5034	*	*

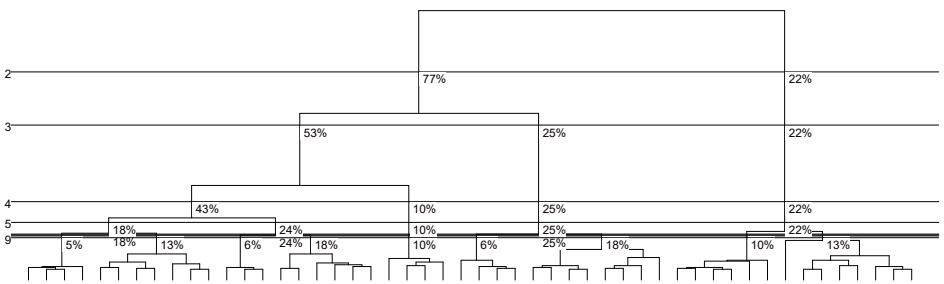
Fonte: elaboração própria.

Nota: por questões de espaço, são omitidas aqui as informações da variância interna e do tamanho das partições com 10 e 50 classes.

Quanto maior o número de classes nas partições, maior é o valor da **variância interclasse** e, portanto, mais elevado é o valor do *eta* quadrado; em contrapartida, diminui a variância intraclasse, pois os agrupamentos, mais numerosos, tornam-se menores e, portanto, internamente mais homogêneos. Notemos também que o ganho adicional na variância explicada com a contínua subdivisão das classes vai decrescendo: assim, quando se passa da partição com 2 classes para aquela com 3 classes, há um ganho na variância explicada (aumento do *eta* quadrado) de 48,9%; quando se passa da partição com 3 classes para aquela com 4, obtém-se um ganho de 17,8%; quando se passa da partição com 4 para aquela com 5 classes, um ganho de 9,44% e assim por diante. É importante saber se esse incremento da variância explicada é acompanhado de ganhos em termos heurísticos.

A observação do dendograma (ou árvore hierárquica) nos apresenta outros aspectos que devem ser considerados na decisão sobre qual partição escolher.

Figura 4.1 | Dendograma exibindo nove partições



Fonte: elaboração própria.

Partindo da parte superior para a inferior do dendograma, devemos observar a distância entre os "nós" resultantes da junção de duas classes. Vemos, por exemplo, que um dos *clusters* se forma em uma etapa anterior de agregação; enquanto outros, apenas em etapas posteriores. É possível notar também que a partição com duas classes (as linhas verticais) está relativamente bem distante da partição com três classes, que também

está bem distante da partição com 4 classes. Esta, por sua vez, está mais próxima da partição com cinco classes, também próxima daquela com seis (as linhas verticais que conectam uma partição à outra são mais curtas). A linha (horizontal) mais grossa é, na verdade, a sobreposição de quatro linhas, resultantes da partição do dendograma em seis, sete, oito e nove classes. De fato, entre a partição com seis classes e aquela com sete classes, há um incremento de pouco mais de 6% na variância explicada, o que é evidenciado pela quase sobreposição das linhas horizontais que cortam o dendograma em partições com seis ou sete classes. Com base na análise dos incrementos na variância explicada quando se passa de uma partição com n classes para outra com n+1 classes e na observação do dendograma, conclui-se que a escolha deve recair em uma partição com quatro ou cinco agrupamentos. A escolha será por uma partição com **cinco agrupamentos**.

Uma vez tendo feito a escolha da solução com "n" *clusters*, o próximo passo é interpretá-los. A interpretação de um agrupamento é feita com base na leitura das estatísticas descritivas calculadas para as variáveis ativas e suplementares, sejam elas numéricas ou categóricas. Dizemos que uma categoria caracteriza um agrupamento se for relativamente mais ou menos frequente no agrupamento do que na amostra; uma variável numérica caracteriza um agrupamento se a média no agrupamento for (significativamente) diferente do que na amostra. Hjellbrekke (2019, p. 85) define três critérios para avaliar se uma categoria está sobrerrepresentada ou sub-representada em um *cluster*: i) se a diferença entre a frequência relativa no agrupamento e na amostra for maior do que 5%; ii) no caso das categorias pouco frequentes, se a frequência no *cluster* for duas vezes maior do que na amostra; e iii) que também tenha um valor *p* menor do que 0,05.

Um modo mais econômico é utilizar a estatística do **valor-teste**, que pode ser usado tanto para variáveis numéricas quanto para variáveis categóricas: se o valor-teste de uma variável numérica for elevado em um agrupamento, significa que ela o caracteriza; no caso de uma variável categórica, se o valor-teste de uma categoria for elevado (geralmente maior do que 1,96, o que equivale a uma probabilidade igual a 0,025 para um teste unilateral), significa que a categoria está sobrerrepresentada

(se o valor for superior a 2) ou sub-representada (se menor do que -2). A seguir, apresentamos as categorias que caracterizam os diferentes agrupamentos[2], que podemos entender como uma **tipologia de práticas culturais ou de consumo cultural.**

Ao fazermos a descrição dos agrupamentos, é importante nomeá-los considerando algumas de suas características distintivas e, então, descrevê-los levando em conta as principais modalidades ativas *e* suplementares com valores-teste superiores ou inferiores a 2.

Tabela 4.2 | Categorias descritivas do primeiro agrupamento

Categorias (ativas e suplementares)	% da categoria no grupo	% da categoria na amostra	% do grupo na categoria	Valor-teste
doc-	73,45	33,10	60,13	16,02
mpb-	84,81	46,34	49,58	15,06
samba-	86,17	54,10	43,15	12,66
espetáculo-	91,38	67,94	36,44	10,19
cinema-	94,06	73,28	34,77	9,67
fundamental incompleto (respondente)	37,54	23,84	42,67	5,86
casado	68,41	56,98	32,53	4,42
trabalhador agrícola (origem de classe)	11,42	6,37	48,55	3,57
½ a 1 salário mínimo (renda domiciliar total per capita)	22,42	15,82	38,40	3,23
entre 35 e 44 anos	28,84	22,35	34,96	2,80

Fonte: elaboração própria.

[2] Como as variáveis categóricas ativas utilizadas são, em sua maioria, dicotômicas, não será necessário observar as duas categorias da mesma variável em dado agrupamento, pois se uma estiver sobrerrepresentada, a outra estará sub-representada (considerando que não há dados ausentes).

O agrupamento dos "**culturalmente desengajados**" abrange 27,09% da amostra, sendo o maior entre os cinco. É assim nomeado porque, nele, estão sobrerrepresentadas as categorias que indicam ausência de consumo cultural ou de preferências por qualquer um dos gêneros de televisão ou de música mencionados (por exemplo: 94,06% dos indivíduos no agrupamento responderam não terem ido ao cinema nos seis meses anteriores à pesquisa em comparação com 73,28% na amostra; ou, então, 84,81% no agrupamento mencionaram não gostar de MPB contra 46,13% na amostra), com exceção da preferência por programas religiosos (ver a Tabela 4.2).[3] A proporção dos que mencionam gostar desse tipo de programa de televisão é ligeiramente maior no agrupamento do que na amostra (48,99% contra 44,02%). Quanto às variáveis suplementares, notemos que estão sobrerrepresentados os indivíduos com baixa escolaridade (fundamental incompleto), trabalhadores manuais não qualificados por conta própria com origem em famílias formadas por trabalhadores rurais, com idade entre 35 e 44 anos, casados e vivendo em domicílios com renda de até 1 salário-mínimo *per capita*.

[3] Por questões de espaço, são apresentadas apenas algumas categorias caracterizadoras de cada agrupamento. Para cada uma delas, há a informação do valor-teste, da proporção da categoria no grupo, da proporção da categoria na amostra e da proporção do grupo na categoria (nesse caso, a proporção dos que "pertencem" a um agrupamento entre os que "escolheram" determinada categoria; por exemplo, das 323 pessoas que responderam que não gostam de documentários, pouco mais de 60% "pertence" ao primeiro agrupamento).

Tabela 4.3 | Categorias descritivas do segundo agrupamento:

Categorias (ativas e suplementares)	% da categoria no grupo	% da categoria na amostra	% do grupo na categoria	Valor-teste
exposição-	98,80	68,37	31,01	12,61
monumento-	84,78	52,35	34,75	10,93
sertaneja+	93,16	67,95	29,42	9,68
livro0	63,14	38,23	35,45	8,20
variedades+	86,76	64,10	29,05	8,00
entre 55 e 64 anos	17,32	8,65	42,99	4,53
casado	70,09	56,98	26,40	4,38
fundamental incompleto (respondente)	34,87	23,84	31,39	4,05
menos do que 1/2 (renda domiciliar total per capita)	19,86	11,68	36,47	3,93
trabalhador manual não qualificado (posição de classe do respondente)	16,34	10,77	32,56	2,66

Fonte: elaboração própria.

O agrupamento dos "**consumidores de cultura popular**" corresponde a quase 21,5% da amostra. Enquanto as categorias que indicam a fruição de bens associados à "alta cultura" (exemplo: leitura de livros, ida a exposições, monumentos, espetáculos, centros culturais e bibliotecas ou, mesmo, ao cinema ou a shows de música ao vivo; ou a rejeição à música erudita, ao jazz ou ao rock) estão sub-representadas (apenas 1,2% dos indivíduos no agrupamento mencionaram terem ido a uma exposição nos doze meses anteriores à pesquisa contra 31,63% na amostra), são relativamente mais frequentes as preferências pelo sertanejo, samba, pagode, música romântica e religiosa, e também por programas de variedades, novelas, musicais e religiosas. Em termos das

variáveis suplementares, destacam-se os indivíduos mais velhos (com 55 anos ou mais), com ensino fundamental incompleto, casados (ou viúvos), trabalhadores manuais não qualificados vivendo em domicílios com, no máximo, ½ salário-mínimo *per capita*.

Tabela 4.4 | Categorias descritivas do terceiro agrupamento

Categorias (ativas e suplementares)	% da categoria no grupo	% da categoria na amostra	% do grupo na categoria	Valor-teste
rap+	89,69	14,55	66,88	19,18
funk+	86,44	13,29	70,57	18,92
pagode+	79,93	42,18	20,55	8,14
rock+	68,83	33,96	21,98	7,69
esporte+	74,20	48,81	16,49	5,37
entre 18 e 24 anos	35,88	19,00	20,48	4,26
trabalhador manual não qualificado (posição de classe do respondente)	20,85	10,77	20,99	3,10
médio completo (escolaridade do respondente)	23,50	13,08	19,50	3,05
menos do que 1/2 sm (renda domiciliar total per capita)	22,15	11,68	20,56	3,01
solteiro	45,49	31,67	15,58	3,00

Fonte: elaboração própria.

O agrupamento dos "**consumidores de cultura contemporânea**" corresponde a quase 11% da amostra. As categorias que mais fortemente o caracterizam são os gostos por *rap*, *funk,* pagode e, em menor medida, por *rock* e por samba (respectivamente, 89,69%, 86,44%, 79,93%, 68,83% e 70,27% contra 14,54%, 13,28%, 42,17%, 33,96% e 45,9% na amostra), e, entre os programas de televisão, por esporte, humor, filmes e variedades. Há também a evitação da leitura e da ida a exposições ou a bibliotecas. Em termos das variáveis suplementares, há um protagonismo dos mais jovens (18 a 34 anos), do sexo masculino, com baixa renda domiciliar,

escolaridade média e com origem em meios populares (em famílias de trabalhadores manuais qualificados ou não).

Tabela 4.5 | Categorias descritivas do quarto agrupamento

Categorias (ativas e suplementares)	% da categoria no grupo	% da categoria na amostra	% do grupo na categoria	Valor-teste
show+	75,07	46,91	30,15	8,51
mpb+	79,36	53,66	27,86	7,90
humor+	81,57	59,15	25,98	7,05
romântica+	87,64	68,17	24,22	6,57
livro1-3	59,04	39,05	28,49	5,89
médio completo (escolaridade do respondente)	21,83	13,08	31,45	3,59
fundamental completo (escolaridade máxima do pai e/ou da mãe)	28,25	18,76	28,36	3,42
entre 3 e 5 sm (renda domiciliar total per capita)	10,71	6,21	32,49	2,63
superior incompleto (escolaridade do respondente)	7,54	3,91	36,34	2,50
supervisor de trabalho manual ou trabalhador manual qualificado (posição de classe do respondente)	21,40	14,88	27,10	2,48

Fonte: elaboração própria.

O quarto agrupamento é o dos "**onívoros culturais**" (18,8% da amostra), cujos repertórios de práticas mais se aproximam de um padrão caracterizado pela mistura de gêneros culturais diferencialmente classificados. Algumas categorias que indicam maior participação cultural caracterizam esse agrupamento: idas a shows, a bares com música ao vivo, a monumentos e, em menor medida, a espetáculos e a exposições, além da leitura de um a três livros nos últimos seis meses. Em termos das preferências musicais,

notemos, aqui, a sobrerrepresentação dos que mencionam gostar de MPB, música clássica, romântica, sertanejo, *jazz*, samba e pagode, além da evitação do *rap* e do *funk*; no que se refere aos programas de televisão, são relativamente mais frequentes os que gostam de documentários, programas de humor, musicais, jornais, filmes e variedades. Quanto às variáveis suplementares, observa-se um protagonismo dos indivíduos com média escolaridade e renda (3 a 5 salários-mínimos), jovens (25 a 34 anos) e com pertencimentos de classe bem delimitados: entre os quadros médios (sendo sua proporção no agrupamento mais do que duas vezes maior em relação à proporção amostral) e entre as frações mais qualificadas dos trabalhadores manuais, com origem em meios populares (pai e/ou mão com fundamental completo).

Tabela 4.6 | Categorias descritivas do quinto agrupamento

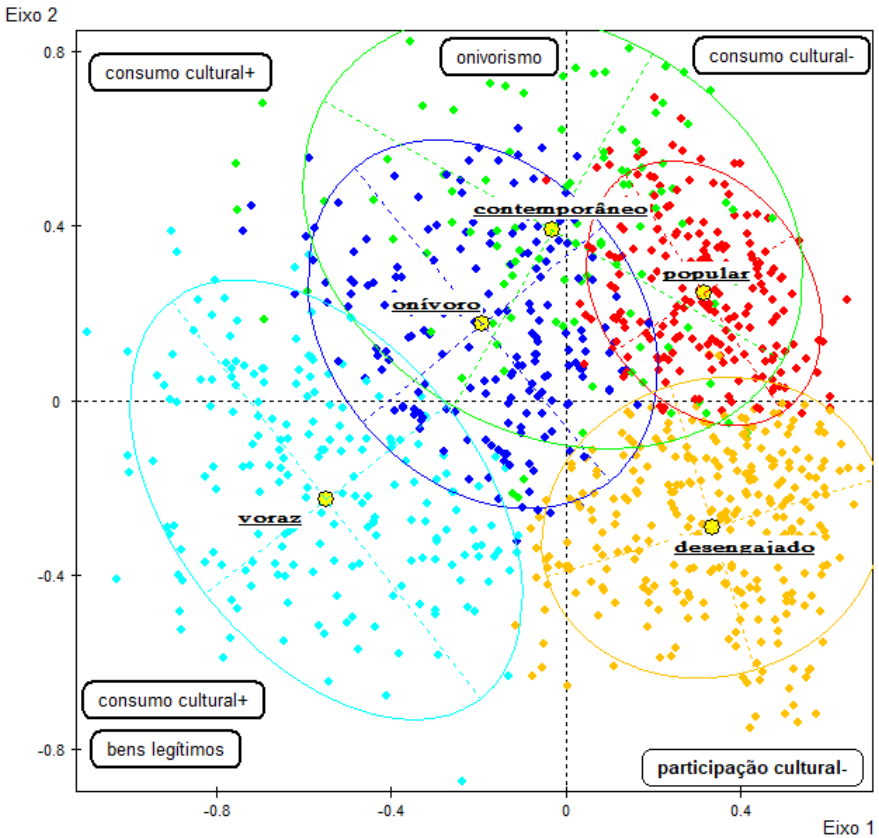
Categorias (ativas e suplementares)	% da categoria no grupo	% da categoria na amostra	% do grupo na categoria	Valor-teste
exposição+	87,02	31,63	59,88	19,40
cinema+	73,17	26,71	59,60	16,42
espetáculo+	76,71	32,06	52,08	15,38
livro4+	61,10	22,73	58,50	14,15
religiosos-	87,34	55,85	34,03	10,92
superior completo (escolaridade do respondente)	35,06	9,49	80,39	12,71
profissionais (posição de classe do respondente)	20,76	6,25	72,24	8,63
superior completo (escolaridade máxima do pai e/ou da mãe)	16,44	5,38	66,48	7,15
solteiro	52,57	31,67	36,12	7,03
5 sm ou mais (renda domiciliar total per capita)	11,49	3,51	71,29	6,05
Entre 18 e 24 anos	31,25	19,00	35,81	4,79

Fonte: elaboração própria.

Por fim, temos o agrupamento dos "**consumidores da alta cultura**" (21,8% da amostra), cujos repertórios de práticas incluem a ida a exposições, cinema, espetáculos, centros culturais, monumentos, bibliotecas, concertos, shows e bares de música ao vivo, além da leitura frequente de livros; as preferências pelo *rock*, por MPB, por documentários e por um conjunto de evitações: sertanejo, pagode, música e programas religiosos, variedades e novelas. Quanto às variáveis suplementares, há maior presença de indivíduos com ensino superior, com renda *per capita* acima de 5 salários-mínimos, com origem social elevada (pai e/ou mãe com ensino superior), e com pertencimentos de classe entre os profissionais e técnicos. Os mais jovens e solteiros são também proporcionalmente mais numerosos aqui do que na amostra em geral.

Uma vez feita a descrição dos agrupamentos, é importante projetá-los no plano fatorial (considerando os eixos retidos para interpretação) como **subnuvens** (resultantes da partição da nuvem global), possibilitando observar as coordenadas de seus pontos médios, assim como a excentricidade e o formato de suas elipses de concentração.

Figura 4.2 | Projeção das elipses de concentração e das subnuvens dos agrupamentos no plano fatorial formado pelos eixos 1 e 2

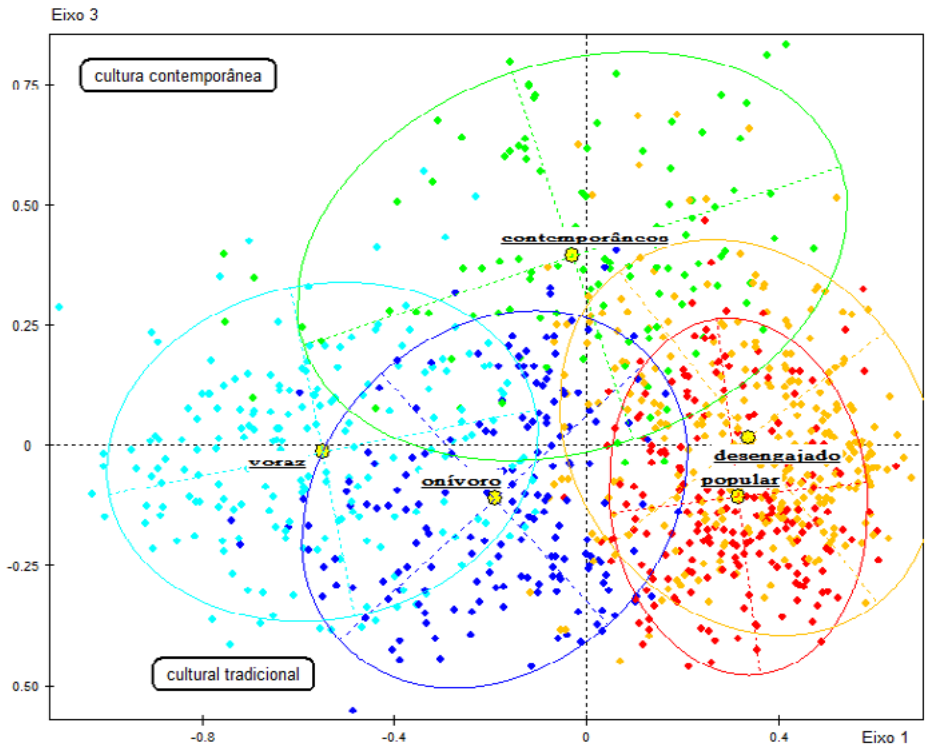


Fonte: elaboração própria.

No plano fatorial formado pelos eixos 1 e 2, a elipse de concentração dos "consumidores de alta cultura" ("voraz") encontra-se quase totalmente no quadrante inferior à esquerda (onde estão as categorias que indicam elevado consumo cultural, marcado pela apropriação de bens da "alta cultura" e rejeição a bens "comuns"); do lado oposto, no primeiro eixo, estão as elipses de concentração dos "desengajados" (quase inteiramente no quadrante onde estão localizadas as categorias que indicam ausência de "participação cultural", tanto no que se refere ao consumo quanto à própria expressão de gostos culturais) e dos

“populares” (no quadrante superior à direita), que diferem dos anteriores no que se refere ao gosto por gêneros musicais e televisivos “ilegítimos”. Diferentemente, a elipse de concentração dos “onívoros” encontra-se a meio caminho entre os dois extremos do primeiro eixo e do segundo eixo (sobrepondo-se, parcialmente, às elipses anteriormente descritas e quase completamente à dos “contemporâneos”). Ocupa uma porção maior à esquerda do primeiro eixo (o lado do consumo cultural elevado). A coordenada do ponto médio da subnuvem no eixo vertical (0,17) é muito similar àquela dos “populares” (0,24), mas seus pontos se espalham mais fortemente para a parte inferior do eixo do que aqueles da subnuvem do agrupamento “popular”. Por sua vez, a elipse que concentra os pontos da subnuvem formada pelos “contemporâneos” ocupa porções similares à esquerda e à direita do primeiro eixo, indicando forte heterogeneidade no que se refere às oposições do eixo horizontal. Diferentemente, encontra-se quase completamente na parte superior do segundo eixo, indicando maior homogeneidade em termos da rejeição à “cultura erudita”.

Figura 4.3 | Projeção das elipses de concentração e das subnuvens dos agrupamentos no plano fatorial formado pelos eixos 1 e 3



Fonte: elaboração própria.

Observando agora o plano fatorial formado pelos eixos 1 e 3, notamos que as principais oposições estão entre os "contemporâneos", cujo ponto médio está localizado bem acima do centro do plano (0,39), de um lado, e os "onívoros" e "populares", de outro, cujos pontos médios estão localizados na parte inferior do eixo (respectivamente, -0,11 e -0,10). Evidências adicionais das oposições entre os agrupamentos no plano fatorial podem ser obtidas pela interpretação do formato e da área das elipses de concentração: a elipse dos "onívoros" ocupa uma porção maior no quadrante inferior à esquerda e a dos "populares" está quase completamente restrita ao quadrante inferior à direita; diferentemente, a elipse dos "contemporâneos" ocupa quase completamente a parte

superior do terceiro eixo e, de forma quase idêntica, os dois lados do eixo 1 (ou seja, indicando, simultaneamente, heterogeneidade em relação a esse eixo e homogeneidade em relação ao eixo 3). As elipses dos "desengajados" e dos "vorazes", embora ocupando porções opostas no primeiro eixo, não se distinguem quanto à localização no terceiro eixo.

Em suma, a CHA confirmou os principais resultados obtidos por meio da ACM nos exemplos aqui tratados: as oposições nos principais eixos anteriormente descritos (participação *versus* não participação cultural; "onivorismo *versus* "seletividade"; "gosto clássico" *versus* "gosto contemporâneo") apareceram aqui novamente, materializadas nas características dos diferentes agrupamentos e em suas localizações relativas nos planos fatoriais.

Conclusão

A combinação da análise de correspondências múltiplas com a análise de classificação hierárquica possibilita, como vimos, classificar os indivíduos e produzir tipologias. O uso das técnicas de classificação permite identificar indutivamente grupos que sejam relativamente homogêneos internamente e diferentes dos demais. Tal estratégia se revela bastante útil quando o objetivo da pesquisa incidir na delimitação de grupos a partir da observação empírica da combinação típica de certas características dos indivíduos apreendidas por meio das modalidades das variáveis ativas e suplementares.

Capítulo 5

Como construir espaços relacionais usando a ACM?

Neste capítulo, gostaria de retomar alguns pontos discutidos anteriormente, considerando as relações entre as dimensões teórica, metodológica e empírica da pesquisa científica. Além disso, gostaria também de abordar algumas das dificuldades resultantes da construção de espaços com a ACM a partir de bases de dados secundários.

Construir um espaço social ou um espaço simbólico nunca se reduz à correta utilização de uma técnica estatística qualquer. Ou seja, dominar o uso da ACM não é condição suficiente para a construção de nuvens (de modalidades e de indivíduos) que nos permitirão responder a determinadas questões de pesquisa. É necessário considerar,

antes de mais nada, o ajuste entre o(s) enquadramento(s) teórico(s), o problema de pesquisa, as hipóteses e os dados (os que produziremos e o que já temos à disposição). Embora não exista qualquer correspondência unívoca entre teoria e metodologia (Platt, 1986; Pires, 2008), no sentido de que determinado enquadramento teórico determine as possíveis escolhas operacionais (exemplo: a técnica de produção dos dados ou o método de análise), não é descabido afirmar que existem algumas "afinidades" entre, de um lado, abordagens teóricas e os problemas de pesquisa construídos a partir delas, e os procedimentos para a coleta e análise dos dados, de outro.

Diferentemente de outras técnicas multivariadas "convencionais" (como a análise de regressão), que possibilitam isolar os efeitos das variáveis "independentes" sobre uma ou mais variáveis "dependentes", a ACM (e, mais amplamente, a AGD) serve essencialmente ao propósito de "explorar e visualizar relações complexas entre variáveis [e suas categorias]... que estão 'escondidas' nos dados" (Roose, 2016, p. 174).

É muito comum associar **o uso da ACM a uma concepção relacional das posições dos agentes e de suas tomadas de posição**: "aqueles que conhecem os princípios da análise de correspondências múltiplas apreenderão as afinidades entre esse método de análise matemática e o pensamento em termos de campo" (Bourdieu, 2001, p. 70; *apud* Lebaron, 2009, p. 13). De forma similar à lógica do conceito de campo, que nos força a "pensar em oposições" (de modo que o sentido e o valor de uma prática sempre devem ser pensados em relação ao sentido e valor de outras), a ACM possibilita considerar as práticas investigadas não em si mesmas, mas em relação umas com as outras, relações essas que podem ser apreendidas por meio, como vimos, das representações gráficas nas nuvens de modalidades e de indivíduos: as distâncias relativas revelando afinidades ou contrastes nas práticas, gostos ou capitais possuídos pelos agentes. O essencial é que o sentido e o valor de uma prática (exemplo: ouvir *rock*) só podem ser apreendidos em relação com o sentido e valor de outras práticas (exemplos: ouvir música clássica, praticar *yoga*, estudar no exterior etc.) e em relação com as posições dos agentes em termos da distribuição dos capitais (exemplo: maior ou menor volume de capital cultural ou econômico). Ou seja, "a ACM é um método relacional à la De

Saussure: as atividades culturais não são consideradas *per se*, mas em relação com outras atividades culturais – dentro do campo de práticas, objetos e disposições" (Roose, 2016, p. 175).

Um argumento similar é levantado por Magne Flemmen em relação aos estudos de classe que se baseiam no conceito de espaço social e na noção a ele associada de **multidimensionalidade**, resultantes das diversas fontes e formas de poder na sociedade. Como argumenta o autor, "essa técnica... permite ao pesquisador construir o espaço social baseado nos indicadores de capital e, assim, proceder de baixo para cima, das distribuições observadas para a dimensionalidade empírica do espaço social" (Flemmen, 2013, p. 329).

Parece trivial afirmar que a escolha de uma técnica (como a ACM) depende das questões que se quer responder com a pesquisa, mas nunca é demais salientar que as decisões aparentemente mais tecnicamente fundamentadas sempre respondem a determinadas problemáticas teóricas.

Além desse "ajuste" entre enquadramento teórico, metodologia e empiria, a construção adequada de espaços com a ACM depende da qualidade dos dados à disposição (sejam eles primário ou secundários). Sabemos, a partir do que foi exposto, que a ACM se aplica a uma base de dados num formato **indivíduos por variáveis**, resultante tipicamente da aplicação de um questionário estruturado a uma amostra de uma população qualquer (é possível também produzir uma base desse tipo a partir da construção de variáveis e categorias derivadas da análise de dados qualitativos, embora existam limitações importantes, especialmente a quantidade de casos). Havendo a possibilidade (tempo, equipe, recursos) para a produção de dados primários por questionário estruturado, é possível já incorporar, no próprio desenho desse instrumento, dois princípios essenciais (Le Roux; Rouanet, 2004, p. 10):

i. o **princípio da homogeneidade** na construção e codificação das variáveis: é importante que as variáveis permitam mensurar diferenças qualitativas, e não apenas quantitativas, no objeto investigado (ou seja, que tenhamos um bom número de variáveis categóricas **nominais**). Por exemplo, em uma pesquisa sobre práticas culturais, além de perguntar a frequência com que um

indivíduo vai ao cinema (variável categórica ordinal), deveríamos perguntar os filmes (e gêneros) que mais gosta e que menos gosta, os lugares a que costuma ir para assistir aos filmes (*shopping center*, cinema de rua etc.), os nomes dos diretores ou das diretoras de cinema que mais aprecia e que menos aprecia e assim por diante. A ACM é uma técnica muito útil para captar tais diferenciações qualitativas nos objetos investigados;

ii. o **princípio da exaustividade**: construção de indicadores que permitam mensurar diversas dimensões ou aspectos do objeto sob investigação. Se o objetivo for o de investigar as práticas culturais, então é importante abordar não apenas o consumo cultural, mas também o gosto e o conhecimento, e não apenas em termos de música, televisão, ou artes, mas também práticas em outros domínios entendidos como culturais em um sentido mais alargado do termo (vestuário, alimentação, educação, cuidados corporais etc.). É só assim que a ACM poderá ser utilizada naquilo em que é mais vantajosa: apreender as relações entre um conjunto muito grande de variáveis e suas categorias (no exemplo em discussão: revelar os padrões complexos que estruturam as práticas, em seus diversos aspectos, dos agentes em inúmeros domínios da vida social).

Além desses princípios, é importante seguir alguns procedimentos para a "correta" utilização da ACM. Todos esses procedimentos foram discutidos anteriormente, por isso, apenas os indicarei brevemente a seguir:

i. **equilibrar o número de variáveis por tópicos e o número de categorias por variável**, com o objetivo de minimizar o risco de que uma variável (ou um conjunto delas) tenha contribuição muito elevada para a inércia total em decorrência simplesmente de definições operacionais;

ii. **evitar incluir como ativas categorias com baixa frequência relativa** (menor ou igual a 5%): nesses casos, a variável pode ser recodificada ou a categoria ser inserida como passiva;

iii. **caso se queira investigar as relações entre diferentes tipos de variáveis (comportamentais, atitudinais/valorativas, sociodemográficas)**, então se um tipo for inserido como ativo, o outro (ou outros) deve(m) ser inserido(s) como suplementar(es);

iv. **analisar o estatuto teórico e distribuição amostral das não respostas** ("não sei"; "não respondeu"): em alguns casos, é possível que a taxa de não resposta seja um dado sociologicamente pertinente; se não for, as categorias das variáveis ativas que indicam resposta ausente devem ser inseridas como passivas. Além disso, deve-se também atentar para a distribuição dos dados ausentes na amostra: alguns indivíduos deixaram de responder a questões mais do que outros? Em que magnitude? Embora não exista uma recomendação única na literatura, é prudente não considerar para a construção das nuvens os indivíduos com respostas ausentes em mais de 20% das questões ativas (Hjellbrekke, 2019, p. 96).

Quando trabalhamos com dados primários, é possível incorporar esses princípios e preocupações no próprio desenho dos instrumentos de coleta de dados. No entanto, é muito comum na pesquisa em ciências sociais utilizarmos bases de dados secundários, de forma subsidiária à produção de dados primários ou como principal fonte de material empírico. Se o uso de dados secundários traz vantagens para a realização da pesquisa (baixo custo, rápida disponibilidade etc.), impõe algumas limitações importantes, sendo a principal delas a de que os dados não foram produzidos conforme a problemática que orienta a pesquisa ou considerando as especificidades das técnicas e métodos de análise dos dados que mais se ajustam aos enquadramentos teóricos mobilizados para a construção do objeto. Por isso, dificilmente teremos todos os indicadores necessários para mensurar os diferentes aspectos pertinentes do objeto (princípio da exaustividade) ou as variáveis necessárias para apreender as diferenciações qualitativas nos casos (princípio da homogeneidade). Será também mais difícil chegar a um equilíbrio do número de variáveis por tópico ou de categorias por variável.[4]

A seguir, gostaria de trabalhar mais dois exemplos de construção de espaços a partir de dados secundários: em um deles, buscarei operacionalizar a noção de espaço social; no outro, de espaço político. Em ambos os casos, o objetivo é mostrar **formatos comuns de nuvens** que obtemos quando trabalhamos especialmente com dados secundários.

[4] Isso não quer dizer, obviamente, que não é possível utilizar dados secundários para a construção de espaços. De fato, os exercícios apresentados anteriormente como exemplos da aplicação da ACM (de suas vertentes) e da CHA se basearam em dados secundários.

A noção de espaço social é central aos estudos de classe de inspiração bourdieusiana. Trata-se de "um espaço multidimensional de posições relativas", diferenciadas conforme a distribuição dos diferentes tipos de propriedades que operam como capitais em uma sociedade. No caso das sociedades capitalistas, tais capitais são, sobretudo, o capital econômico (recursos patrimoniais, fundiários, monetários etc.); o capital cultural (conjuntos de saberes em estado incorporado, institucionalizado ou objetivado); e o capital social (relações a que os agentes podem recorrer para satisfazer seus interesses). Tais capitais se distribuem desigualmente em termos de volume e composição (mais de um tipo ou menos de um tipo na estrutura global) e conforme as modalidades de apropriação (trajetórias sociais). Tal noção de espaço social é de difícil operacionalização, pois supõe a construção de um conjunto grande de indicadores para os diferentes tipos de capital, a partir de um corte sincrônico e um diacrônico. Caso queiramos operacionalizar uma noção desse tipo utilizando bases secundárias, precisamos buscar nelas variáveis que possam funcionar como indicadores de capital econômico, cultural e social (minimamente dos dois primeiros tipos) e que permitam apreender diferentes trajetórias de acumulação, conservação ou transformação da estrutura do capital (indicadores de origem social).

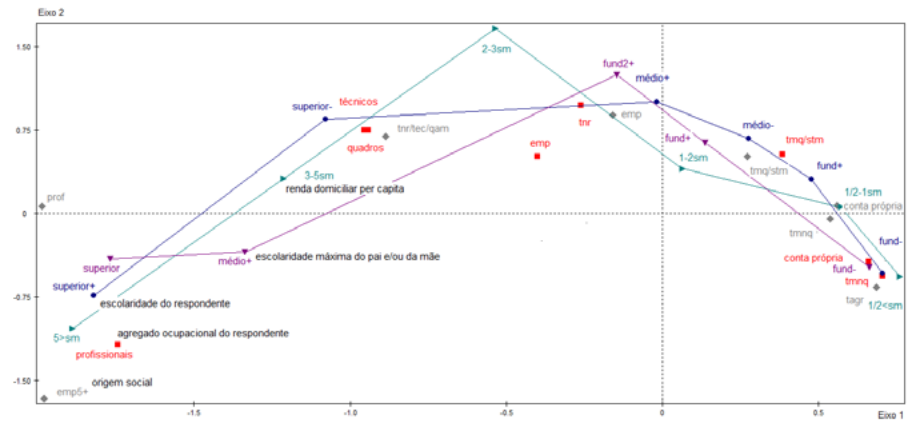
Utilizarei a mesma base de dados dos exercícios anteriores para operacionalizar esse conceito de espaço social.[5] Serão utilizadas 5 variáveis: renda domiciliar *per capita* (6 categorias), escolaridade do respondente (7 categorias), escolaridade máxima do pai e/ou da mãe (5 categorias), agregado ocupacional do respondente (8 categorias), agregado ocupacional do pai ou da mãe (8 categorias, sendo três delas passivas).

Observando a Figura 5.1, a seguir, vemos que o padrão de distribuição das modalidades se assemelha a uma "ferradura": nesse caso, o primeiro eixo opõe os valores extremos das categorias, enquanto o segundo eixo opõe aqueles aos valores médios. A interpretação do segundo eixo não é possível a menos que consideremos os contrastes no primeiro eixo. Formatos desse tipo refletem a **unidimensionalidade** na distribuição

[5] Aguiar, Neuma *et al*. Pesquisa por amostragem sobre desigualdades sociais na região metropolitana de Belo Horizonte, 2005 (Banco de Dados). Belo Horizonte: UFMG, 2005. In: Consórcio de Informações Sociais. 2017. Disponível em: <http://www.nadd.prp.usp.br/cis/index.aspx>. Acesso em 11/09/2020.

das categorias das variáveis, resultado que, em grande medida, tem a ver com os tipos de variáveis utilizadas: como todas elas são variáveis categóricas de tipo ordinal (inclusive, em certa medida, as variáveis de agrupamento ocupacional), é possível produzir apenas diferenciações quantitativas, e não qualitativas. Obviamente, poderíamos chegar a resultados similares mesmo se tivéssemos a possibilidade de produzir diferenciações sociologicamente mais sutis. No entanto – e este é o ponto – o uso de variáveis ordinais na ACM tende a produzir comumente nuvens com o referido formato de "ferradura". A nuvem de indivíduos, por sua vez, tende a assumir um formato triangular.[6]

Figura 5.1 | Projeção das categorias ativas no plano fatorial formado pelos eixos 1 e 2

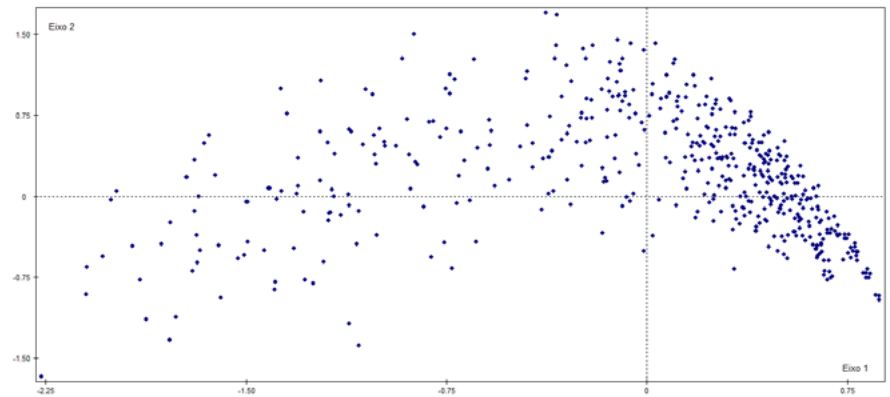


Fonte: elaboração própria.

[6] Além da recodificação das variáveis, uma estratégia para lidar com esse problema consiste em incluir cada variável duas vezes. Como explica Hjellbrekke (2019, p. 96), uma variável mensurada em uma escala que vai de 1 a 5 é recodificada em duas variáveis com escalas, respectivamente, de 0 a 4 e de 4 a 0. Na literatura, encontramos outras maneiras de lidar com esse problema, que, na verdade, tem a ver com a conversão de variáveis métricas em "falsas" variáveis categóricas (Cf. Le Roux; Rouanet, 2004, p. 219-221; Greenacre, 2017; Aşan; Santurk, 2011).

Quando chegamos a resultados desse tipo, devemos considerar, para a interpretação, o plano fatorial, e não cada eixo separadamente. Além disso, é importante considerar as oposições nos demais eixos, além do primeiro e do segundo (que tendem a ter forte contribuição para a inércia total).

Figura 5.2 | Nuvem de indivíduos projetada no plano fatorial formado pelos eixos 1 e 2



Fonte: elaboração própria.

No exemplo a seguir, será feita uma tentativa de operacionalizar o conceito de espaço político, que busca apreender as relações entre tomadas de posição quanto a questões políticas. Com base em trabalhos recentes sobre o tema (Flemmen; Haakestad, 2018; Harrits *et al*. 2010), utilizarei como variáveis ativas cinco questões sobre justiça distributiva. Outras questões poderiam ser usadas, mas, para os propósitos deste exercício, essas bastarão. São elas:

i. "É fácil para crianças que vêm de família pobre terem boa educação" (Q1).

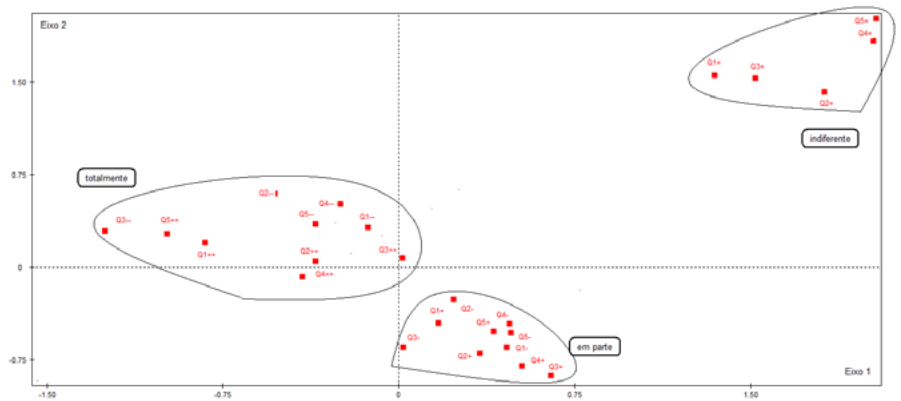
ii. "O valor da aposentadoria que o governo paga para os idosos deveria aumentar mesmo que isso signifique que pessoas como você tenham que pagar maiores impostos" (Q2).

iii. "Pessoas como você pagam impostos demais" (Q3).

iv. “As pessoas pobres deveriam receber do governo uma renda mínima em torno de 200,00 reais por mês” (Q4).

v. “As pessoas pobres têm a possibilidade de melhorar de vida (sair da pobreza) no Brasil hoje” (Q5).[7]

Figura 5.3 | Projeção das categorias ativas no plano fatorial formado pelos eixos 1 e 2



Fonte: elaboração própria.

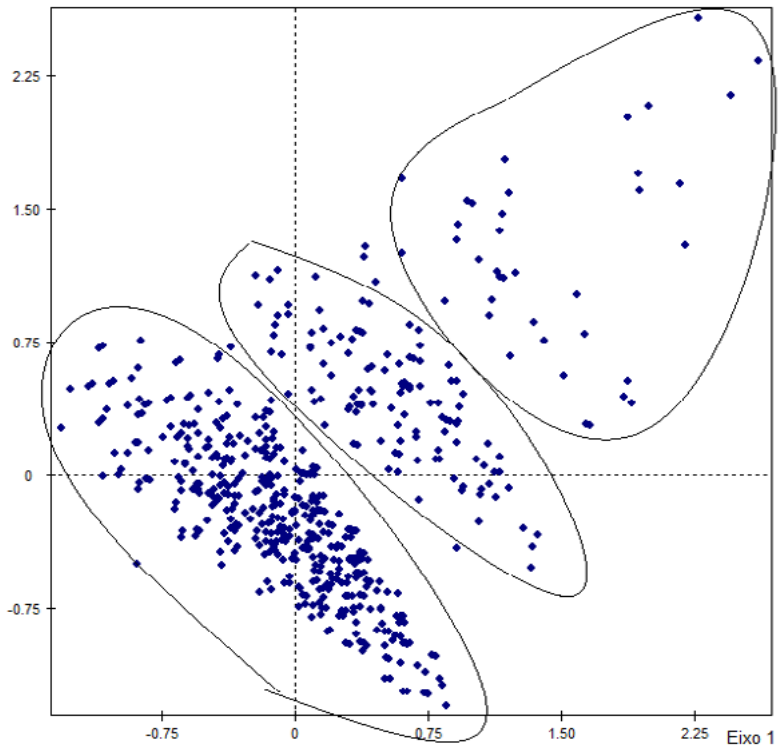
Observando a nuvem de categorias (Figura 5.3), notamos que as modalidades que indicam “indiferença” (nem concorda nem discorda) estão concentradas à direita do primeiro eixo e acima do segundo; aquelas que indicam concordância ou discordância parcial, também à direita do primeiro eixo, mas abaixo do segundo; e, por fim, as modalidades que indicam concordância ou discordância total estão quase todas concentradas à esquerda do primeiro eixo e acima do segundo. Tal padrão de distribuição das categorias no plano fatorial é o resultado

[7] Os indivíduos poderiam responder escolhendo as seguintes categorias de resposta: “concordo”, “discordo” ou “nem concordo nem discordo”. Caso respondessem “concordo” ou “discordo”, então, deveriam dizer se “em parte” ou “totalmente”, somando, então, cinco categorias de resposta.

Aguiar, Neuma *et al*. Pesquisa por amostragem sobre desigualdades sociais na região metropolitana de Belo Horizonte, 2005 (Banco de Dados). Belo Horizonte: UFMG, 2005. In: Consórcio de Informações Sociais, 2017. Disponível em: <http://www.nadd.prp.usp.br/cis/index.aspx>. Acesso em: 11/09/2020.

provavelmente de algum nível de redundância entre as categorias ativas (Hjellbrekke, 2019, p. 98). No exemplo anterior, é provável que o mesmo subconjunto de indivíduos tenha escolhido principalmente a categoria de resposta "nem concorda nem discorda" para as questões. Outro subconjunto escolheu as categorias extremas (positivas ou negativas) para todas ou quase todas, e assim por diante. De fato, quando observamos a nuvem de indivíduos (Figura 5.4), percebemos um padrão desse tipo.[8]

Figura 5.4 | Projeção da nuvem de indivíduos no plano fatorial formado pelos eixos 1 e 2



Fonte: elaboração própria.

[8] Uma maneira de contornar esse problema é por meio da recodificação das variáveis. Juntando as respostas "em parte" e "totalmente", ficamos com três categorias de respostas para cada questão: "concorda" (em parte ou totalmente), "discorda" (em parte ou totalmente) e "nem concorda nem discorda". Seria importante também considerar o significado da resposta "indiferente". É possível que a resposta "nem concorda nem discorda" seja mais a expressão de uma dificuldade de produzir uma opinião do que propriamente o resultado de uma reflexão quanto à problemática trazida pela questão. Tal entendimento nos levaria, então, a inserir tal modalidade como passiva.

Conforme argumenta Di Franco (2016), os formatos mais comuns das nuvens de modalidades e de indivíduos são os seguintes:

i. **elipses**, que ocorrem quando os primeiros eixos são os fatores mais importantes na variância total;

ii. **nuvens separadas**, "quando dois subconjuntos não se interpenetram", como no exemplo anterior, é possível fazer análises separadas;

iii. **ferradura**, "configuração [que] indica uma certa unidimensionalidade na distribuição das frequências empíricas porque o segundo fator representado é interpretável apenas em termos do primeiro" (Di Franco, 2016, p. 1306).

A vantagem de trabalharmos com dados primários, quando usamos a ACM (ou qualquer outra das técnicas da AGD), é que podemos construí-los considerando as peculiaridades dessa técnica: no caso da ACM, devemos considerar os dois princípios básicos antes mencionados, de modo que tenhamos um conjunto de indicadores que nos permitirão apreender diferenciações qualitativas no objeto e classificar nossos casos por meio de um conjunto de variáveis categóricas nominais (e não simplesmente ordinais). Dessa forma, teremos melhores resultados. Obviamente, isso não significa que a ACM não possa ser proveitosamente empregada para análises de dados secundários, como, aliás, os exemplos anteriores – espero – bem demonstraram, desde que saibamos dos limites e possibilidades da referida técnica e como usá-la adequadamente na pesquisa empírica.

Bibliografia

Agresti, Alan; Finlay, Barbara. *Métodos estatísticos para as ciências sociais*. 4. ed. Porto Alegre: Penso, 2012.

Aşan, Zerrin; Santurk, Sevil. An Application of Fuzzy Coding in Multiple Correspondence Analysis for Transforming Data from Continuous to Categorical. *Journal of Multiple-valued Logic and Soft Computing*, v.17, n. 5-6, p. 591-606, 2011.

Bennett, Tony *et al*. *Culture, class, distinction*. Londres / Nova Iorque: Routledge, 2009.

Benzècri, Jean-Paul. *Correspondence analysis handbook*. Nova Iorque: Dekker, 1992.

Benzècri, Jean-Paul *et al*. *l'analyse des données*. Paris: Dunod, 1973.

Bertoncelo, Edison. Classe social e alimentação: padrões de consumo alimentar no Brasil contemporâneo. *Revista Brasileira de Ciências Sociais*, v. 34, n. 100, p. 1-27, 2019a.

_____. Consumo cultural e manutenção das distâncias sociais no Brasil. In: Pulici, Carolina; Fernandes, Dmitri. *As lógicas sociais do gosto*. São Paulo: Edunifesp, 2019b.

_____. O espaço das classes sociais no Brasil. *Tempo Social*, v. 28, n. 2, p. 73-104, 2016.

_____. O uso da análise de correspondências múltiplas nas ciências sociais: possibilidades de aplicação e exemplos empíricos. In: Encontro Anual da Anpocs, 40., 2016, Caxambu - MG. *Anais*... Caxambu - MG: Anpocs, 2016. Disponível em: https://www.anpocs.com/index.php/encontros/papers/40-encontro-anual-da-anpocs/st-10/st16-7/10296-o-uso-da-analise-de-correspondencias-multiplas-nas-ciencias-sociais-possibilidades-de-aplicacao-e-exemplos-empiricos/file.

Blasius, Jörg; Greenacre, Michael. *Visualization and verbalization of data*. Boca Raton: CRC Press, 2014.

Bourdieu, Pierre. *A distinção*: crítica social do julgamento. São Paulo: Edusp; Porto Alegre: Zouk, 2008.

_____. *Science de la science et réflexivité*. Paris: Raisons d'agir, 2001.

Cano, Ignacio. Nas trincheiras do método: o ensino da metodologia das ciências sociais no Brasil. *Sociologias*, v. 14, n. 31, p. 94-119, 2012.

Clausen, Sten-Erik. *Applied correspondence analysis*: an introduction. Thousand Oaks: Sage Publications, 1998.

Di Franco, Giovanni. Multiple correspondence analysis: one only or several techniques? *Quality & Quantity*, v. 50, p. 1299-1315, 2016.

Flemmen, Magne. Putting Bourdieu to work for class analysis: reflections on some recent contributions. *The British Journal of Sociology*, v. 64, n. 2, p. 325-343, 2013.

Flemmen, Magne; Haakestad, Hedda. Class and politics in twenty-first century Norway: a homology of positions and position-taking. *European Societies*, v. 20, n. 3, p. 401-423, 2018.

Greenacre, Michael. *Correspondence analysis in practice*. Londres / Nova Iorque: Chapman & Hall/CRC, 2017.

Greenacre, Michael; Blasius, Jörg. *Multiple correspondence analysis and related methods*. Londres / Nova Iorque: Chapman & Hall / CRC, 2006.

Harrits, Gitte S.; Prieur, Annick; Rosenland, Lennart; Sjott-Larsen, Jakob. Class and politics in Denmark: are both old and new politics structured by class? *Scandinavian Political Studies*, v. 33, n. 1, p. 1-27, 2010.

Hjellbrekke, Johs. *Multiple correspondence analysis for the social science*. Abingdon / Nova Iorque: Routledge, 2019.

Husson, François; Josse, Julie. Multiple correspondence analysis. In: Blasius, Jörg; Greenacre, Michael. *Visualization and verbalization of data*. Boca Raton: CRC Press, 2014.

Klüger, Elisa. Análise de correspondências múltiplas: fundamentos, elaboração e interpretação. *Revista Brasileira de Informação Bibliográfica em Ciências Sociais,* v. 86, n. 2, p. 68-97, 2018.

Lahire, Bernard. Campo. In: Catani, Afrânio M.; Nogueira, Maria Alice; Hey, Ana Paula; Medeiros, Cristina. *Vocabulário Bourdieu*. Belo Horizonte: Autêntica, 2017.

Lebaron, Frédéric. How Bourdieu 'quantified' Bourdieu: the geometric modelling of data. In: Robson, Karen; Sanders, Chris. *Quantifying theory*: Pierre Bourdieu. Dordrecht: Springer Science, 2009.

Lebaron, Frédéric; Bonnet, Philippe. Class specific analysis: methodological and sociological reflections. In: Blasius, Jörg *et al*. (org.) *Empirical investigations of social space*. Gewerbestrasse: Springer, 2019.

Lebart, Ludovic *et al*. *Statistique exploratoire multidimensionelle*. Paris: Dunod, 1998.

Lebart, Ludovic; Saporta, Gilbert. Historical elements of correspondence analysis and multiple correspondence analysis. In: Blasius, Jörg; Greenacre, Michael. *Visualization and verbalization of data*. Boca Raton: CRC Press, 2014.

Le Roux, Brigitte. Structured data analysis. In: Blasius, Jörg; Greenacre, Michael. *Visualization and verbalization of data*. Boca Raton: CRC Press, 2014.

Le Roux, Brigitte; Rouanet, Henry. *Multiple correspondence analysis*. Thousand Oaks: Sage, 2010.

_____. *Geometric data analysis*: from correspondence to structured data analysis. Dordrecht: Kluwer Academic Publishers, 2004.

Nogueira, Cláudio Marques Martins. Espaço social. In: Catani, Afrânio M.; Nogueira, Maria Alice; Hey, Ana Paula; Medeiros, Cristina. *Vocabulário Bourdieu*. Belo Horizonte: Autêntica, 2017.

Pereira, José Virgílio B. P. *Classes e culturas de classe na cidade do Porto*: classes sociais e "modalidades de estilização da vida" na cidade do Porto. Porto: Edições Afrontamentos, 2005.

Perosa, Graziela Serroni; Lebaron, Frédéric; Leite, Cristina Kerches da Silva. O espaço das desigualdades educativas no município de São Paulo. *Pro-Posições*, v. 26, n. 2, p. 99-118, 2015.

Peterson, Richard. Problems in comparative research: the example of omnivorousness. *Poetics*, v. 33, n. 5-6, p. 272-282, 2005.

Platt, Jeniffer. Functionalism and the survey: the relation of theory and method. *The Sociological Review*, v. 34, n. 3, p. 501-536, 1986.

Pires, Alvaro. Sobre algumas questões epistemológicas de uma metodologia geral para as ciências sociais. In: Poupart, Jean *et al*. *A pesquisa qualitativa*: enfoques epistemológicos e metodológicos. Petrópolis: Editora Vozes, 2008.

Rouanet, Henry. The geometric analysis of structured individuals x variables tables. In: Greenacre, Michael; Blasius, Jörg. *Multiple correspondence analysis and related methods*. Boca Raton / Londres / Nova Iorque: Chapman & Hall / CRC, 2006.

Roose, Henk. Getting beyond surface: using geometric data analysis in cultural sociology. In: Hanquinet, Laurie; Savage, Mike. *Routledge international handbook of the sociology or art and culture*. Londres / Nova Iorque: Routledge, 2016.

Sayer, Andrew. *Method in social science*: a realist approach. Londres: Sage, 1992.

Savage, Mike. The musical field. *Cultural Trends*, v. 15, n. 2/3, p. 159-174, 2011.

Schroeder, Larry D.; Sjoquist, David L.; Stephan, Paula E. *Understanding regressions analysis*. Beverly Hills: Sage Publications, 1986.

Silva, Glauco Peres. *Desenhos de pesquisa*. Brasília: Enap, 2018.