

**Tipologias Institucionais e Análises Quantitativas sobre a  
Educação Superior**

***Institutional Typologies and Quantitative Analyses of Higher  
Education***

***Tipologías Institucionales y Análisis Cuantitativas sobre la  
Educación Superior***

Santos, Renato Augusto dos <sup>(i)</sup>

Raposo, Bruno Marão <sup>(ii)</sup>

Carrijo, Maria Clara Silva Cesar <sup>(iii)</sup>

<sup>(i)</sup> Instituto Nacional de Estudos e Pesquisas Educacionais Anísio – INEP, Diretoria de Avaliação da Educação Superior – DAES, Brasília, DF, Brasil. Universidade Federal do Rio de Janeiro – UFRJ, Programa de Pós-Graduação em Sociologia e Antropologia – PPGSA, Rio de Janeiro, RJ, Brasil. <https://orcid.org/0009-0001-5909-3024>, [br.rsan@gmail.com](mailto:br.rsan@gmail.com).

<sup>(ii)</sup> Universidade Federal do Rio de Janeiro – UFRJ, Programa de Pós-Graduação em Sociologia e Antropologia – PPGSA, Rio de Janeiro, RJ, Brasil. Universidade de Brasília, UnB – Brasília, DF, Brasil. <https://orcid.org/0009-0005-3662-5525>, [brunoraposo2@gmail.com](mailto:brunoraposo2@gmail.com).

<sup>(iii)</sup> Universidade Federal do Rio de Janeiro – UFRJ, Programa de Pós-Graduação em Sociologia e Antropologia – PPGSA, Rio de Janeiro, RJ, Brasil. Universidade de Brasília, UnB – Brasília, DF, Brasil. <https://orcid.org/0009-0009-6143-1123>, [clara.carrijo@gmail.com](mailto:clara.carrijo@gmail.com).

## Resumo

A produção de estudos científicos ou subsidiários à formulação e à implementação de políticas públicas em cada Educação Superior (ES) deve contar com informações da sua própria dinâmica interna. Usualmente, essas análises tomam como referência recortes definidos a partir das categorizações previstas na estrutura formal da ES, que não são capazes de direcionar de forma unívoca as decisões estratégicas e pedagógicas de mantenedores e gestores das Instituições de Educação Superior (IES). Nessa direção, aborda-se a aplicação de métodos estatísticos como outra forma de aproximação das dinâmicas da ES, permitindo um olhar que transponha a estrutura prevista na legislação vigente, ao desvelarem-se recortes de análise delimitados a partir dos dados representativos da realidade das IES.

**Palavras-chave:** Tipologia Institucional, Educação Superior, Análises Multivariadas, Análise Fatorial, Análise de Clusters.

## Abstract

*The production of scientific studies or supporting research for the formulation and implementation of public policies in each Higher Education (HE) should rely on information derived from its own internal dynamics. Typically, these analyses are based on categorizations defined for the formal structure of the HE, which are not capable of unequivocally guiding the strategic and pedagogical decisions of HEI managers and maintainers. In this regard, the application of statistical methods is considered as an alternative approach to understanding the dynamics of an HE, offering a perspective that transcends the structure outlined in current legislation by revealing analytical categories defined based on data that truly represent the reality of HEIs.*

**Keywords:** Institutional Typology, Higher Education, Multivariate Analyses, Factor Analysis, Cluster Analysis.

## Resumen

*La producción de estudios científicos o que respalden a la formulación e implementación de políticas públicas en cada Educación Superior (ES) debe basarse en información sobre su propia dinámica interna. Habitualmente, estos análisis se apoyan en recortes definidos según las categorizaciones definidas para la estructura formal de la ES, que no orientan de forma unívoca las decisiones estratégicas y pedagógicas de los gestores y entidades mantenedoras de las Instituciones de Educación Superior (IES). En este sentido, se plantea la aplicación de métodos estadísticos como otra forma de aproximarse a las dinámicas de la ES, permitiendo una mirada que trascienda la estructura legal vigente, al revelar recortes analíticos delimitados por datos representativos de la realidad de las IES.*

**Palabras clave:** Tipología Institucional, Educación Superior, Análisis Multivariantes, Análisis Factorial, Análisis de Clústeres.

## Introdução

O presente texto tem como escopo o delineamento metodológico adotado na pesquisa intitulada “Medindo a relação entre a diversidade institucional e equidade estudantil em países da América Latina”<sup>1</sup>, mais especificamente em relação à parte destinada a identificar tipos institucionais a partir da aplicação de métodos estatísticos, tendo como recorte o funcionamento das Instituições de Educação Superior (IES) da educação superior (ES) de cada país – Argentina, Brasil, Chile, Peru e Uruguai.

São apresentadas as decisões tomadas na aplicação dos métodos estatísticos de análises multivariadas (AM) sobre os dados atinentes ao funcionamento das IES da ES brasileira, abrangendo as etapas de aplicação das AM até a análise dos dados, com vistas a apresentar o delineamento da metodologia quantitativa aplicada na pesquisa, com as devidas caracterizações dos métodos utilizados e suas aplicações no processo de produção do conhecimento em andamento.

Para definir que dados deveriam compor o rol de variáveis a serem utilizadas nas análises estatísticas, foi importante manter em vista o objetivo geral do estudo: avaliar em que medida diferentes arranjos institucionais e tipos de instituições que surgiram dentro dos sistemas de ES, com o resultado de expansão e diversificação, têm sido capazes de aumentar a equidade de oportunidades educacionais de acesso e participação de determinados grupos sociais historicamente desfavorecidos – “grupos de equidade”.

Do objetivo definido, derivam duas principais ações, que refletem uma estratégia comum em pesquisas que analisam as diferentes dinâmicas de expansão e inclusão na ES em todo o mundo: 1) a criação de tipologia institucional que permita analisar a organização das IES e 2) o seu uso para analisar como diferentes sistemas educacionais nacionais se expandiram e que impactos produziram sobre as desigualdades sociais, por meio da inclusão de estudantes provenientes de grupos historicamente desfavorecidos.

<sup>1</sup> Esse projeto de pesquisa é desenvolvido no âmbito do Centro Latino-Americano de Pesquisa em Educação Superior (CeLapes), instalado no Colégio Brasileiro de Altos Estudos (CBAE), e do Laboratório de Pesquisa em Ensino Superior (Lapes), vinculado ao Instituto de Filosofia e Ciências Sociais (IFCS) da Universidade Federal do Rio de Janeiro (UFRJ), mantendo interação com outros de seus projetos. Como financiadores dos projetos encontram-se as seguintes entidades: Society for Research into Higher Education (SRHE), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

Esse recorte adotado mantém relação com temas de pesquisa específicos sobre tipologias e ensino superior (Huisman et al., 2015; Teixeira et al., 2013; Van Vught, 2009), desigualdades sociais e ensino superior (Alon, 2009; Lucas & Moore, 2001;), e no uso de tipologias para analisar a estratificação no ensino superior (Barbosa & Santos, 2011; Croxford & Raffe, 2014; Fumasoli & Huisman, 2013). Já a abordagem metodológica sustentada por métodos quantitativos foi inspirada em estudos de membros do grupo de pesquisa sobre a educação superior (Rodrigues, 2022; Vieira, 2021), que se inspiraram, por sua vez, em outros trabalhos acadêmicos com abordagens semelhantes.

## **Tipologia institucional definida a partir de dados empíricos**

O uso de categorizações em análises envolvendo a ES de um país se torna indispensável quando as quantidades de variáveis e unidades de observação inviabilizam um olhar, geral e abrangente sobre o conjunto de cursos e instituições desse nível educacional, fazendo-se necessários sistematizações e agrupamentos de informações, como no caso do Brasil. Nesses cenários, a categorização é algo recorrente e indispensável no âmbito de estudos operacionais e acadêmicos ou da operacionalização de política pública, em todo o seu ciclo de vida.

Normalmente, essas categorizações tomam como referência a estrutura formal e alguns aspectos operacionais previstos na legislação vigente, o que tem se mostrado uma abordagem bastante relevante, mas que pode dificultar a identificação de determinadas dinâmicas presentes no ES. Como consequência, sob a perspectiva da atuação do Poder Público, pode-se levar algum tempo para que pontos de atenção sejam identificados e políticas públicas preventivas ou remediativas, com objeto de intervenção bem delimitado, sejam propostas, mesmo estando-se diante de um contexto político favorável.

Uma tipologia institucional sobre as IES brasileiras, apoiada na categorização oficial, pode ser observado no estudo “Por uma tipologia do ensino superior brasileiro: teste de conceito” (Schwartzman et al., 2021), que toma as categorizações oficiais atinentes às naturezas administrativas das mantenedoras e aos tipos de organização acadêmica das IES como ponto de partida para a apresentação de uma tipologia institucional. A partir das categorizações oficiais, foram evidenciadas características secundárias para estabelecer subcategorias ou subgrupos de

IES, fazendo uso de dados relativos aos cursos de graduação, programas de pós-graduação *stricto sensu* – de mestrado e de doutorado – e a quantidade de matrículas para definir o porte da IES.

Não se pode negar que os aspectos da estrutura formal no Brasil apontam para fatores que tendem a exercer influência na forma como as IES se comportam no Sistema Nacional de Educação Superior (SNES). Entretanto, conforme estudos e a experiência têm mostrado, os tipos de mantenedora e de organização acadêmica das IES não são capazes de fazer com que essas atuem de forma equivalente na ES e alcancem resultados semelhantes nos processos formativos de seus estudantes.

Essas diferenças de atuação ocorrem porque as disposições normativas não estabelecem limites restritivos à atuação das IES, mas somente requisitos mínimos a serem cumpridos; e concedem autonomia de gestão interna para as IES definirem suas estratégias de atuação, atendendo às demandas sociais circundantes e aos próprios interesses institucionais ou de suas mantenedoras. Com isso, não direcionam de forma unívoca as decisões estratégicas e pedagógicas de mantenedores e gestores das IES.

Efeitos da autonomia das mantenedoras sobre as definições de estratégias de atuação para suas IES podem ser encontrados no texto intitulado “A formação de oligopólios da educação superior brasileira” (Zuccarelli et al., no prelo), onde se encontram evidências de que a estrutura formal relativa aos tipos de IES previstos não determinam de forma unívoca as escolhas estratégicas sobre a atuação das IES do país.

Diferentes formas de atuação podem ser identificadas, também, quando se comparam as IES que possuem categorizações formais comuns, como, por exemplo, as universidades privadas com fins lucrativos; ou, ainda, comparações envolvendo outros recortes definidos a partir de categorizações oficiais. É justamente por essa razão que estudos sustentados por análises estatísticas podem viabilizar a identificação de dinâmicas da ES que se apresentem em grupos diferentes daqueles estabelecidos na legislação vigente.

Essa forma de aproximação do objeto de estudo tem potencial de evidenciar aspectos sobre o fenômeno que podem ter sua relevância atenuada ou mesmo negligenciada nas análises que têm por foco identificar os aspectos de semelhança e de diferenciação entre as IES, dado o olhar dos pesquisadores permeado por suas experiências pessoais em relação ao objeto de estudo (Chizzotti, 2003). Sob essa perspectiva é que esse eixo da pesquisa é realizado, com o delineamento metodológico apresentado na seção subsequente deste texto, buscando-se

identificar novas formas de categorização das IES para além das tipologias institucionais definidas na legislação. Com isso, os dados utilizados nas análises estatísticas são relativos aos aspectos de funcionamento das IES da ES.

Certamente, não se pretende dizer que a utilização da estrutura formal ou dos aspectos operacionais descritos na legislação perde sua relevância nos processos de análise sobre a ES de um país ou qual abordagem em si é melhor que outra. O que se pretende evidenciar é que existem técnicas de análise que são mais adequadas a determinado problema de pesquisa, com determinado recorte, ou a determinada análise que vislumbra novas formas de aproximação da realidade para identificar novas nuances de análise.

## **Aspectos metodológicos das análises quantitativas**

A abordagem quantitativa em estudos toma a realidade como algo objetivo e mensurável, buscando quantificar fenômenos e utilizar métodos estatísticos para testar hipóteses, geralmente formuladas antes da coleta de dados numéricos, que ocorre por meio de instrumentos apropriados (Chizzotti, 2003). Essa abordagem também permite a testagem de teorias emergentes, a criação de novas hipóteses teóricas e o desenvolvimento de novas teorias (Creswell, 2010; Glaser & Strauss, 1967; Yin, 2016).

Ainda, permite identificar resultados que demonstrem padrões ou correlações condizentes ou não com teorias existentes, evidenciando anomalias que contrariem as teorias consolidadas ou emergentes (Creswell, 2010; Glaser & Strauss, 1967; Yin, 2016). Diante disso, surge a necessidade de serem produzidas novas teorizações para explicar essa nova nuance do fenômeno em análise (Yin, 2016). Seu emprego pelo pesquisador também pode ocorrer em situações complexas ou processos em que não se tem clareza sobre as variáveis relevantes (Yin, 2016).

Em pesquisas alicerçadas em métodos quantitativos, também são aplicados métodos qualitativos na fase exploratória para se obter a base conceitual, evidenciadora das categorias de análise sobre as quais os métodos quantitativos são estruturados (Luna, 1998). Isso se torna ainda mais relevante em contextos complexos e multifacetados (Yin, 2016), como o enfocado na pesquisa aqui explicitada. Ainda, permite ao pesquisador identificar variáveis relevantes e

desenvolver hipóteses, assim como construir instrumentos de coleta de dados mais apropriados às subsequentes análises estatísticas (Chizzotti, 2003; Creswell, 2010).

O processo de definição de uma tipologia institucional para a ES brasileira resultante dessa pesquisa, baseada nos aspectos de funcionamento das IES, está sustentado em métodos estatísticos de análise multivariada (AM), que servem para medir, explicar e prever o grau de relação entre variáveis estatísticas, que devem ser aleatórias e inter-relacionadas de forma que seus efeitos não possam ser interpretados separadamente (Hair et al., 2009).

Para viabilizar a análise e a interpretação da grande quantidade de variáveis que compõem o banco de dados constituído para fins de caracterizar aspectos de funcionamento das IES, foram escolhidos dois métodos estatísticos de AM, com vistas a objetivos específicos e distintos, embora com certo ponto de interação, buscando-se identificar com a:

- a) Análise Fatorial (AF) – dimensões de comportamento da ES, evidenciados pelos grupos de características de funcionamento das IES encontrados a partir das correlações existentes entre as variáveis, que podem servir de mediadoras na interpretação dos grupos de IES identificados com métodos de agrupamento.
- b) Análise de Clusters (AC) – os grupos de IES que apresentem aspectos de funcionamento semelhantes, que permitam identificar uma tipologia institucional delimitada a partir desses atributos das IES.

A aplicação desses métodos é possível por meio de cálculo computacional, fazendo-se uso de *softwares* que permitem realizar análises estatísticas robustas. No âmbito da pesquisa foram utilizados o *software* “R” e os pacotes estatísticos “psych”, “mclust” e “tidyLPA”. O pacote “psych” está sendo utilizado na análise fatorial. Já para a análise de clusters, está em uso o pacote “Mclust”, que apresentava melhor desempenho no processo de cálculo e mais opções de modelos a serem testados e disponibilizados, quando comparado com o pacote “tidyLPA”.

Esses métodos trabalham com algoritmos aleatórios, gerando variáveis randômicas não correlacionadas, o que pode gerar resultados diferentes cada vez que o método é reaplicado para o mesmo conjunto de dados. Para inibir esse comportamento, é indicado definir uma semente de aleatoriedade na programação para replicar os mesmos resultados toda vez que o cálculo for refeito.

Toda aplicação do método estatístico passa por diversas fases, que podem ser elencadas e descritas de diferentes formas. Com base em alguns autores (Crespo, 2022; Triola, 2013) e em experiências em pesquisas quantitativas, entende-se possível elencar as seguintes: definição do problema de pesquisa ou hipóteses a serem testadas, planejamento do desenvolvimento da pesquisa, coleta de dados, tratamento dos dados, análise dos dados, interpretação dos resultados e apresentação dos resultados.

O delineamento aqui apresentado, cobrindo o desenvolvimento da pesquisa desde a fase de seu planejamento até a análise dos dados por meio de AM, foi resultante de muitas idas e vindas, abarcando revisões das dimensões norteadoras da coleta de dados, das variáveis utilizados no estudo e dos métodos e técnicas estatísticos utilizados – sendo esse último grupo de revisões perpassado pela realização dos cálculos computacionais previstos para as AM, que abarcaram, inclusive, verificações de hipóteses teóricas emanadas da teoria estatística.

Dado o peso dos métodos estatísticos nessa pesquisa, optou-se pela realização de estudo detalhado sobre os métodos e técnicas estatísticas definidas para seu desenvolvimento. Assim sendo, o estudo abrangendo as teorias estatísticas buscou verificar a pertinência de utilização de cada método e técnica de acordo com os interesses, os objetivos e o tipo de dados selecionados para as análises; e estabelecer bases teóricas sólidas em suas aplicações na pesquisa.

## **Planejamento da Abordagem Quantitativa da pesquisa**

A fase de planejamento do desenvolvimento no estudo possibilita organizar o processo de produção de conhecimento e as definições prévias sobre as dimensões ou características mensuráveis do fenômeno, as fontes de dados, os instrumentos que viabilizarão essa coleta dos dados e as técnicas estatísticas que serão utilizadas nas análises.

No contexto da pesquisa aqui apresentada, essa fase mostrou-se importante, pois permitiu antecipar os recursos necessários para o seu desenvolvimento e algumas possíveis dificuldades de operacionalização, além das dimensões do objeto de estudos a serem utilizadas, as fontes secundárias de dados e as análises estatísticas a serem aplicadas, com vistas ao objeto de estudos e aos objetivos da pesquisa.

## Coleta dos dados

A coleta de dados pode ser caracterizada como direta, quando proveniente de fonte primária, ou indireta, quando proveniente de fonte secundária. Na pesquisa, foram utilizadas fontes secundárias, em que a coleta ocorre sobre dados obtidos para outros fins ou em outro estudo correlato, realizado por outro pesquisador ou instituição, por meio da aplicação de instrumentos construídos para atender aos seus objetivos (Crespo, 2002).

A escolha por determinado nível de agregação das variáveis depende dos objetivos do estudo e das análises estatísticas a serem realizadas. Entretanto, quando se utiliza fonte secundária, o ideal é que o pesquisador tenha conhecimento, no momento de definir os objetivos e o recorte da pesquisa, de que dados estão disponíveis e em que nível de agregação são apresentados. Caso contrário, na fase de coleta de dados, o objetivo e o recorte da pesquisa precisam ser ajustados ou, até mesmo, totalmente reformulados.

Nessa perspectiva, o recorte da pesquisa, a abrangência do olhar sobre o objeto de estudos ou o escopo de suas análises estatísticas acabam por ser definidos em decorrência dos dados disponíveis ou dos seus níveis de agregação. Isso ocorreu na pesquisa, pois nem todos os aspectos de funcionamento das IES são cobertos pelos dados coletados ou disponibilizados pelo Governo Federal. Por exemplo, seria desejável encontrar dados em suficiência e qualidade para representar as atividades de extensão e de cursos de pós-graduação *lato sensu* ou de outras naturezas, para além dos cursos de graduação e pós-graduação *stricto sensu*, o que acabou não acontecendo.

Quanto ao nível de agregação dos dados, à época da coleta, os dados foram encontrados nas fontes oficiais: a) em seu menor nível de agregação, como microdados; e b) em diversos níveis de agregação, como no caso de sinopses estatísticas, onde os dados são dispostos segundo as intenções definidas para esse tipo de apresentação dos dados. Então, sob essa perspectiva, a pesquisa não foi afetada pelo nível de agregação dos dados disponíveis na fonte secundária, podendo-se optar pela utilização de microdados.

Os dados coletados de fonte secundária foram provenientes das coletas realizadas via Censo da Educação Superior e via avaliação dos programas de pós-graduação *stricto sensu* – de mestrado e de doutorado. Ambas abrangem todo o SNES e são realizadas, respectivamente, pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP),

anualmente; e pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), quadrienalmente, com atualização anual de informações de matrículas.

Os dados do Censo da Educação Superior de 2019 contavam com dados de matrículas e de docentes no nível de agregação por indivíduo, anonimizados e dispostos de forma a não permitir a reidentificação dos indivíduos sem a aplicação de técnicas de mineração de dados mais sofisticadas. Por outro lado, os dados da avaliação quadrienal mais recentes foram coletados em nível de agregação por programa de pós-graduação *stricto sensu*, com informações de matrículas atualizadas em 2019.

Muitos estudos têm se debruçado sobre a conceitualização e a mensuração da diversidade institucional, enfocando, sobretudo, os sistemas de educação superior (SES) americanos e europeu. Nesse tipo de estudo, a identificação de dimensões relevantes para análise exige uma compreensão conceitual das características importantes das IES, que podem estar relacionadas a considerações tanto teóricas e epistemológicas quanto funcionais dos SES (Huisman et al., 2015).

Para selecionar os dados a serem utilizados nas análises estatísticas foram definidas dimensões do funcionamento das IES que permitissem identificar e agrupar os dados disponíveis. Nos processos de revisão de literatura e de coleta de dados quantitativos, outras dimensões foram consideradas e algumas rearranjadas para melhor refletir o funcionamento da ES no contexto brasileiro e nos demais países.

Nessa direção, a partir da análise da literatura e da legislação vigente, com vistas ao objetivo definido para o estudo, foram definidas seis dimensões para direcionar a seleção dos dados oficiais disponíveis: a) governança, b) ensino, c) pesquisa, d) extensão ou terceira missão, e) internacionalização e f) aspectos caracterizadores de tamanho das IES, que se desdobraram em subcategorias de análise.

## Tratamento dos dados

Realizada a coleta de dados, teve início a fase de tratamento dos dados, quando: i) foi verificada a qualidade de cada variável coletada ou calculada; ii) foram calculadas novas variáveis a partir das coletas, com nível de agregação por IES; iii) foram selecionadas as variáveis calculadas que seriam utilizadas nas AM; e iv) foi realizado reescalonamento das variáveis com vistas à AC.

Sobre o nível de agregação dos dados, embora sempre se tenha pretendido trabalhar no nível de IES, algumas medidas foram constituídas tomando como base aspectos de funcionamento dos cursos de graduação, como, por exemplo, quando se toma como enfoque a modalidade de oferta ou a área de conhecimento do ensino de graduação, segundo a Cine Brasil (Inep, 2019).

A maior parte das variáveis calculadas tomou como base os microdados do Censo da Educação Superior de 2019, expressando aspectos das IES, sempre na direção de representar as três possibilidades ou obrigações das IES em relação ao ensino, à extensão e à pesquisa. Já as variáveis calculadas a partir dos dados da avaliação quadrienal são referentes à oferta de programas de mestrado ou de doutorado, às quantidades de matrículas vinculadas aos programas e às suas produções bibliográficas.

Para a pesquisa, foram calculadas 156 variáveis, dentre as quais somente 60 foram selecionadas para as AM, segundo a última revisão da lista de variáveis realizada pelos pesquisadores. Além de critérios de pertinência, fonte dos dados originais e qualidade das variáveis obtidas, também foi considerado o tipo de variável – quantitativa ou qualitativa. Esse último critério se deve ao elemento complicador que variáveis qualitativas trazem para a operacionalização das AM<sup>2</sup>, sendo todas as variáveis desse tipo deixadas de fora por não terem sido consideradas indispensáveis para representar aspectos de funcionamento da IES.

Ainda houve outras decisões relacionadas à seleção de variáveis, relacionadas a:

---

<sup>2</sup> Embora existam métodos estatísticos para transformar variáveis qualitativas em quantitativas, como o escalonamento ideal, nenhuma aplicação em estudo similar foi indicada na literatura analisada. Com isso, para toda a aplicação de recursos estatísticos no âmbito dessa pesquisa, optou-se pela não utilização de métodos e técnicas para além daqueles previamente testados nos tipos de análise multivariada utilizadas na pesquisa.

- a) Categorizações oficiais previstas para as IES, tendo-se em vista que: i) conforme abordado anteriormente, essas não expressam diretamente a forma como as IES atuam na ES; e ii) a presença dessas categorizações nas AM poderia produzir efeitos indesejados nos agrupamentos das IES, na medida em que podem se tornar variáveis agregadoras das unidades de observação por apresentarem pouca variabilidade.
- b) Indicativos de inclusão de estudantes com perfil socioeconômico historicamente desfavorecido, por terem sido consideradas inadequadas para a definição da tipologia institucional objeto dessa pesquisa, tendo-se em vista o efeito circular que produziriam nas análises subsequentes focadas na relação do funcionamento das IES com a equidade de acesso e participação desses estudantes.

A qualidade das variáveis originais e calculadas foi verificada por meio de análises estatísticas descritivas, buscando identificar as variáveis que não deveriam fazer parte das AM em decorrência de características de seus valores ou da ausência de variabilidade, com o objetivo de mitigar impactos desfavoráveis nas AM, em consonância com recomendações existentes na literatura estatística (Alzahrani et al., 2021; Haynes, Ricaherd & Kubany, como citado em Laros, 2012).

Isso é necessário, porque variáveis que apresentam pouca informação podem gerar ruídos ou problemas tanto para a AF quanto para a AC, uma vez que apresentam baixa variabilidade e, conseqüentemente, baixa correlação com outras variáveis, podendo produzir uma matriz de correlações singular, que tem seu determinante nulo e, portanto, não inversível. Esse foi um objeto de atenção, que culminou na retirada de 1 (uma) das 60 (sessenta) variáveis selecionadas para as AM, referente à proporção de matrículas na área da Cine Brasil denominada “Serviços”.

Essa verificação se deve ao fato de a AF e a AC se utilizarem da matriz de correlação no processo de cálculo, mas com finalidades diferentes: i) a AF usa-a para alocar a variável em um fator, tendo como efeito da pouca informação a ausência de alocação da variável em algum fator ou a geração de ruído no processo de alocação das variáveis.; e ii) a AC usa-a para definir o tipo de modelo a ser utilizado no agrupamento das unidades de observação, considerando a adoção ou não de variabilidade para os parâmetros relativos a volume, formato e orientação, tendo como efeito da pouca informação a configuração de um caso de singularidade e inviabilizar a utilização de determinados modelos.

Ainda com foco na qualidade das variáveis, foram realizadas análises de correlação entre as variáveis calculadas e selecionadas para as AM, fazendo-se uso do método de Pearson por serem todas quantitativas e contínuas. Isso porque a ausência de correlações ou a colinearidade entre variáveis pode gerar uma matriz de correlação singular, o que ocasiona desdobramentos indesejados na AF e na AC.

A escolha do método a ser selecionado para a análise de correlação decorre do tipo de variável selecionada para o estudo a ser desenvolvido, sendo a Pearson, quando se trata de variáveis quantitativas contínuas, e a policórica, no caso de variáveis qualitativas ordinais. Quando há a presença de variáveis quantitativas e qualitativas, é indicada a retirada das qualitativas ou, sendo indispensável sua presença na AM, a sua transformação em variáveis dicotômicas, que representem cada uma das possíveis categorias da variável original, sendo o atributo assumido originalmente indicado na nova variável com o valor 1 (um) e os demais atributos sinalizados com o valor 0 (zero), aplicando-se posteriormente o método de Pearson (Hair et al., 2019).

Outro tratamento de dados a ser considerado é o reescalonamento de variáveis<sup>3</sup>, quando se têm valores exorbitantes entre os diferentes dados a serem utilizados na AC, tendo em vista que escalas muito díspares podem gerar ruído no processo de agrupamento das unidades de observação, ao trazerem diferentes pesos nas estimativas, dificultando a geração dos clusters (Bouvyron et al., 2019). Embora essa necessidade também seja algo associado à AF, essa ideia não tem fundamentação estatística, pois essa transformação não produz desdobramentos sobre as análises de correlação e, por conseguinte, nos resultados obtidos.

Quando o reescalonamento se faz necessário para a AC, desde que não se trate de variável geoespacial, devem-se utilizar, dentre os métodos de reescalonamento de variáveis disponíveis, aqueles que tenham como resultado o desvio padrão das variáveis igual a 1 (um) (Bouvyron et al., 2019). Tendo sido o caso dessa pesquisa, em que havia diferenças muito

<sup>3</sup> Na literatura, é possível encontrar outras formas de transformações de variáveis, como a logarítmica e a exponencial. Mas, conforme explicitado anteriormente, optou-se pela não utilização de métodos e técnicas para além daqueles previamente testados nos tipos de análise multivariada utilizados na pesquisa.

grandes entre as escalas das variáveis selecionadas para realização da AC, tendo sido usado o método de padronização<sup>4</sup> para mitigar eventuais ruídos no processo de geração dos clusters.

Finalizado o tratamento dos dados aqui apresentados, pode-se proceder com a aplicação das AM propriamente ditas.

## Análise fatorial

A Análise Fatorial é um método utilizado para investigar diversas variáveis por meio de seus padrões e suas relações, organizando-as em um conjunto de fatores, que correspondem a um conjunto mínimo de variáveis não observadas diretamente, mas que representam o agrupamento das variáveis originais mais correlatas entre si e mantêm relação direta com as dimensões ou construtos destacados do objeto de estudo (Johnson & Wichern, 2007, Matos & Rodrigues, 2019).

Nessa perspectiva, a AF busca facilitar o estudo de uma grande quantidade de variáveis por meio do redimensionamento de fatores explicativos (Johnson & Wichern, 2007), determinando, inclusive, se essas podem ou não ser redimensionadas. Entretanto, o redimensionamento resultante da AF pode ocasionar perda significativa da variabilidade dos dados, algo indesejado em AM, sendo necessário verificar a efetividade do redimensionamento para representar as variáveis originais em análises subsequentes.

Os fatores são combinações lineares, criadas a partir da matriz de correlação de variáveis originais, produzindo um escore para cada unidade de observação. Para a construção dos fatores, o método associa cargas fatoriais, que variam de  $-1$  a  $1$ , para determinar a alocação de cada variável ao seu respectivo fator, que são consideradas significativas quando seu valor absoluto for igual ou superior a  $0,3^5$ , tornando o fator um representante útil da variável. As cargas fatoriais podem, por sua vez, ser consideradas como pesos de regressão das variáveis mensuradas para predizer o construto subjacente, representando o quanto a variável original está associada a um determinado fator (Laros, 2012).

<sup>4</sup> O método de padronização de variáveis consiste em transformar as medidas originais em valores referenciados na distância entre o valor original e a média de todos os valores da variável, com essa diferença dividida pelo desvio padrão.

<sup>5</sup> Outros valores existentes na literatura podem ser empregados como critérios de análise, mas utilizou-se a indicação do referido autor na aplicação desse método estatístico no âmbito desta pesquisa.

Existem dois tipos de AF, que têm como elementos de diferenciação as hipóteses obtidas *a priori* e os seus objetivos (Laros, 2012; Matos & Rodrigues, 2019), conforme explicitado a seguir:

- a) Análise Fatorial Exploratória (AFE) – usa os dados como referência para definir os fatores (ou dimensões) que agrupam as variáveis, não se baseando em qualquer hipótese *a priori*; trata-se de uma abordagem exploratória, porque o investigador não tem qualquer expectativa firme preexistente sobre os fatores a serem criados.
- b) Análise Fatorial Confirmatória (AFC) – usa os dados para fins de verificar as dimensões ou construtos do objeto de estudos previamente definidos com base em teorias existentes ou na experiência empírica, testando uma hipótese prévia de agrupamento.

De acordo com o tipo de AF em que foi gerada, o valor absoluto da carga fatorial tem interpretação diferente. Na AFE, quando a variável não obtém uma carga fatorial significativa em nenhum dos fatores ou é variável única do fator, tem-se ausência de relação dessa com as demais variáveis, devendo ser excluída do conjunto da análise. Depois, a AFE deve ser refeita com base no subconjunto remanescente, sendo esse procedimento repetido até que todas as variáveis remanescentes sejam alocadas em algum fator (Laros, 2012). Já na AFC, isso significa que a hipótese previamente definida não foi confirmada empiricamente, não cabendo o mesmo procedimento indicado no âmbito da AFE.

A aplicação da AF como forma de redução de variáveis é algo comum nesse tipo de estudo para: i) viabilizar cálculos computacionais subsequentes, quando o excesso de dados leva os *softwares* a não produzirem resultados; ii) atender ao princípio da parcimônia, com vistas a se evitar o ajuste excessivo – *overfitting* – em processos de modelagem sobre uma população a partir de amostras, mitigando-se interferências que o conjunto original de variáveis podem gerar em AM subsequentes; ou iii) viabilizar análises de resultados decorrentes de outras AM que envolvam muitas variáveis<sup>6</sup>.

Nenhum dos casos se aplica a esta pesquisa, tendo-se em vista que: a) a utilização do pacote “Mclust” resolveu o problema de cálculo computacional e está sendo utilizado na análise de clusters o universo das IES do SNES; b) a perda do poder explicativo das variáveis no

<sup>6</sup> A AF também pode ser aplicada para tratar casos de multicolinearidade entre variáveis, quando esse tipo de relação entre as variáveis se torna um problema para o método estatístico em uso, o que não é o caso da AC.

processo de redimensionamento indicou que seria melhor utilizar o conjunto dos dados brutos original; c) a utilização de fatores, apesar da perda no poder explicativo, seria aceitável somente se fosse indispensável o atendimento ao princípio da parcimônia, não sendo o caso desta pesquisa, tendo em vista a utilização da população, e não de amostra; d) a utilização dos fatores, apesar da perda no poder explicativo, com vistas a facilitar as análises dos resultados da AC, na próxima etapa da aplicação da metodologia estatística na pesquisa, não se mostrou um caminho vantajoso, tendo em vista a quantidade elevada de variáveis que não foram agregadas a nenhum dos fatores.

Outra finalidade associada à AFE decorre diretamente das correlações imbricadas na estrutura desse método, que permitem identificar grupos de variáveis que indicam tendências no conjunto das unidades de observação. No âmbito desta pesquisa, a AFE tem essa aplicação, permitindo identificar comportamentos da ES, evidenciados pelos grupos de características de funcionamento das IES destacados a partir das correlações existentes entre as variáveis selecionadas com vistas ao desenvolvimento da pesquisa.

Ainda, as dinâmicas existentes no SNES, identificadas pelos padrões de comportamento similares das variáveis relativas ao funcionamento das IES, expressos pelos fatores, pode servir de recorte de análise e interpretação dos grupos de IES resultantes da AC. Nessa perspectiva, embora não seja uma aplicação com garantia de sucesso, configurando-se como mais um recurso de análise de resultados, a utilização dos fatores pode servir para identificar mais facilmente as similaridades que levaram à formação dos grupos de IES a partir da AC.

De um modo geral, a aplicação da AF em um estudo deve percorrer as seguintes etapas, envolvendo o conjunto de dados selecionado para essa análise: 1) verificação de pré-requisitos e de critérios de adequação do método, 2) definição do tipo de extração dos fatores, 3) definição da quantidade mínima de fatores, 4) realização da AF sem rotação, 5) interpretação das relações entre os fatores criados para fins de definição do tipo de rotação do modelo deve ser aplicado, 6) realização da AF com o tipo de rotação definido e 7) análise dos resultados da AF.

## **Critérios de viabilidade de aplicação do método**

Para a aplicação da AF, tem-se como pré-requisito a existência de correlação entre as variáveis a serem redimensionadas em fatores. As correlações devem orientar o processo de seleção correta das variáveis a serem redimensionadas, indicando, inclusive, aquelas que devem ser retiradas por apresentarem algum tipo de ruído ou não terem correlação significativa.

No tocante à matriz de correlação, ainda cabem dois testes para verificar a viabilidade de aplicação da AF sobre um determinado conjunto de dados: o *Kaiser-Meyer-Olkin* (KMO), também conhecido como índice de adequação da amostra (*measure of sampling adequacy* - MSA) – aplicado em casos de análises envolvendo amostras e população –, e o Teste de Esfericidade de Bartlett (Hongyu, 2018).

O KMO sugere que a proporção da variância dos itens pode ser explicada por uma variável latente ou fator, indicando o quão adequada é a aplicação da AF para o conjunto de dados. Os critérios em relação aos índices de KMO são os seguintes: os valores menores que 0,5 são considerados inaceitáveis, os valores entre 0,5 e 0,7 são considerados medíocres, os valores entre 0,7 e 0,8 são considerados bons e os valores maiores que 0,8 e 0,9 são considerados, respectivamente, ótimos e excelentes<sup>7</sup>. Ainda, a partir do valor geral do KMO é possível calcular um valor para cada variável – o MSA –, que pode ser utilizado como mais um critério de análise, com vistas à retirada de variáveis que apresentam índice de MSA menor que 0,6 para se tentar aumentar o valor geral do KMO (Garson, 2022).

O teste de esfericidade de Bartlett é um método que avalia em que medida a matriz de covariância é similar a uma matriz identidade, com os termos de covariâncias entre as variáveis iguais a zero. A rejeição da hipótese nula indica que a matriz de covariância apresenta valores significativos que podem ser agrupados em fatores e, portanto, o conjunto de variáveis analisado pode ser utilizado para fins de aplicação na AF (Hongyu, 2018).

<sup>7</sup> Outros valores existentes na literatura podem ser empregados como critérios de análise, mas utilizou-se a indicação do referido autor na aplicação desse método estatístico no âmbito desta pesquisa.

## Definição da quantidade de fatores e do método de extração dos fatores

Após verificada a viabilidade de se utilizar a AF para o conjunto de variáveis selecionadas para o estudo, fazem-se necessárias duas definições metodológicas, independentes e não sequenciadas: a quantidade de fatores e o método para a extração dos fatores.

Para determinar a quantidade de fatores necessária, podem ser utilizados métodos baseados nos autovalores de cada fator, sendo a Análise Paralela de Horn considerado o melhor método para essa finalidade (Laros, 2012). Esse método utiliza matrizes de correlação calculadas a partir de variáveis randômicas não correlacionadas, com a mesma dimensão do banco de dados original. A partir dos autovalores da matriz original e das matrizes simuladas, representadas por um valor crítico para cada fator, são apresentadas duas linhas no gráfico do tipo *Scree Plot*, em que o ponto de interseção indica a quantidade de fatores a serem extraídos, sendo a quantidade mais indicada aquela imediatamente anterior ao ponto de interseção das linhas.

Em relação à definição do método de extração dos fatores, existem dois que são mais amplamente utilizados, de acordo com tipo de distribuição dos dados, sendo utilizada: i) a máxima verossimilhança (*Maximum likelihood* - ML) para os dados com distribuição normal, em que a extração é feita a partir de estimação de parâmetros, que são os mais prováveis possíveis de reproduzirem a matriz de correlação observada, quando o conjunto dos dados apresenta uma distribuição normal multivariada (IBM, 2021); e ii) os componentes principais, quando os dados não apresentam distribuição normal (Costello & Osborne, 2005).

Apesar de a Análise de Componentes Principais (ACP) ser um método de redimensionamento de variáveis em si, também pode ser utilizado no processo de extração de fatores (Laros, 2012). Nessa perspectiva, a ACP busca explicar a maior parte da informação original (variância) a um número mínimo de fatores. Logo, a partir da decomposição espectral, extraem-se fatores que explicam de forma decrescente os maiores percentuais da variância, sendo que o primeiro fator explica o maior percentual, o fator subsequente explica o segundo maior percentual e, assim, sucessivamente. Dessa forma, a partir da ACP é possível achar as cargas fatoriais em termos de autovalores-autovetores (Hair et al., 2009).

Para se chegar a essa conclusão, como método de verificação do tipo de distribuição dos dados em cada variável, foi utilizado o Teste de Shapiro-Wilk com nível de significância igual a 5%, tendo sido rejeitada a hipótese nula de normalidade para todas as variáveis (Shapiro & Wilk, 1965).

## Rotação do modelo

No processo de geração dos fatores, é desejado que todos tenham carga fatorial positiva e que as variáveis não fiquem concentradas em poucos fatores para se ter uma estrutura fatorial simples. Quando isso não ocorre, faz-se necessário aplicar a transformação na matriz de fatores e na matriz de cargas fatoriais por meio de rotação (Laros, 2012).

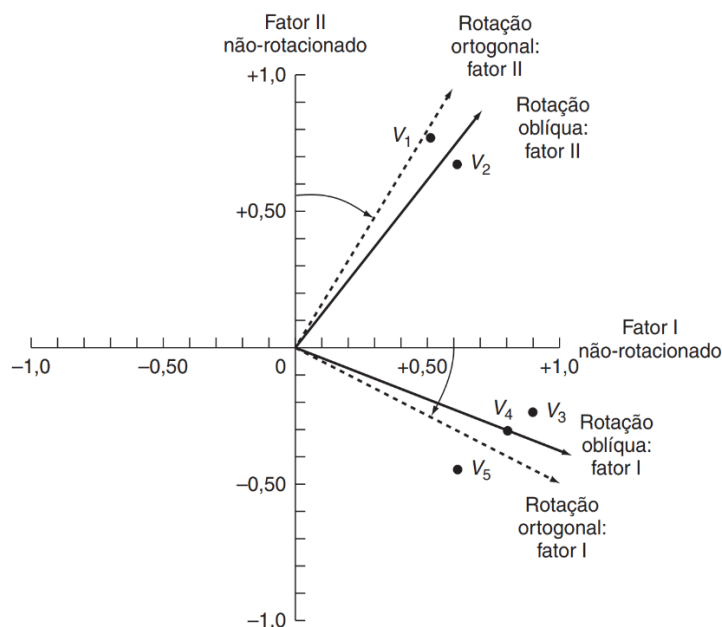
A AF é um método que se assemelha à regressão, sendo a variável resposta equivalente à variável mensurada e as variáveis explicativas equivalentes aos fatores criados, cujas cargas representam os pesos de cada fator que explicam cada variável mensurada. Realizado o mesmo procedimento para todas as variáveis, obtém-se, então, o modelo matricial da AF.

A rotação é realizada por meio da transformação na matriz dos fatores e na matriz das cargas fatoriais, em que se utiliza a matriz identidade, que em termos matemáticos não altera o modelo produzido a partir das matrizes mensuradas (Manly & Alberto, 2017). São utilizadas duas formas de rotação:

- a) Rotação ortogonal – apropriada para modelos fatoriais em que os fatores comuns são considerados independentes, sendo o procedimento mais comum o Varimax – que objetiva maximizar a variância das cargas fatoriais para cada fator por meio do aumento de cargas altas e a diminuição das cargas baixas. Essa maximização consiste em se obter melhor distinção de cada variável associada a cada fator, fazendo com que cada variável tenha uma carga fatorial alta e as demais baixas, evitando a situação indesejada de as variáveis terem múltiplas cargas fatoriais altas em diversos fatores. Este procedimento é a opção padrão em quase todos os pacotes estatísticos e produz uma estrutura fatorial razoavelmente simples na maioria das situações (Laros, 2012).
- b) Rotação oblíqua – apropriada para modelos fatoriais em que os fatores podem ser correlacionados, sendo o procedimento mais comum o Promax, que, dentre os diversos procedimentos disponíveis, conduz a uma estrutura simples mais convincente e melhor interpretável do que uma solução ortogonal (Laros, 2012).
- c) A representação gráfica dessas duas formas de rotação em relação aos eixos do plano cartesiano consta na Figura 1.

**Figura 1**

*Rotação dos eixos do plano cartesiano nas rotações ortogonal e oblíqua*



Fonte: Hair et al. (2009)

## Análise de Clusters

A Análise de Clusters é um tipo de método utilizado para encontrar grupos significativos nas variáveis analisadas, delimitando grupos em que se agregam as unidades de observação mais parecidas, segundo suas características comuns, e se discriminam aquelas mais diferentes, de forma que os grupos de unidades de observação formados sejam homogêneos entre si e heterogêneos diante dos demais (Bouveyron et al., 2019).

Nessa perspectiva, esse método busca somente padrões ou grupos similares, sem conferir atenção ao significado político ou conceitual das variações identificadas nos dados (Betarelli Junior & Ferreira, 2018). Porém, os aspectos conceituais envolvidos no processo de seleção de variáveis acabam se tornando um importante referencial na atribuição de significado aos resultados obtidos com a AC, viabilizando uma melhor caracterização dos clusters delimitados a partir dos dados selecionados para esse fim.

Os tipos de método de AC vão desde aqueles amplamente heurísticos até os que adotam procedimentos mais formais, baseados em modelos estatísticos, seguindo, geralmente, uma estratégia hierárquica ou uma estratégia em que as observações são realocadas entre grupos

provisórios (Fraley & Raftery, 1998). Dentre os diferentes tipos, deve ser selecionado aquele que melhor se aplica ao objeto de estudo e ao conjunto de variáveis selecionadas para as análises (Betarelli Junior & Ferreira, 2018). Não se pode perder de vista que cada método apresenta determinado nível de complexidade envolvido no processo do cálculo computacional e de erros de estimação que perpassam os resultados produzidos.

A aplicação da AC tem um papel importante nesta pesquisa, diante do objetivo de definir uma tipologia institucional para uma determinada ES com bases nos aspectos de funcionamento de suas IES. Com isso, fez-se necessário analisar os tipos de métodos de AC mais usados e identificar o que melhor se adequaria aos objetivos e aos dados selecionados para o desenvolvimento da pesquisa como um todo.

Na literatura foi encontrado um teste envolvendo três tipos de métodos de clusterização para se identificar aquele com menor erro de estimação: *k-means*, *single-link* e *por modelagem*. Foi verificado que nem todos são aplicáveis para fins de clusterização, quando a distribuição das unidades de observação nos clusters apresenta sobreposição ou o formato não é esférico<sup>8</sup>, sendo, nesses casos mais adequado o uso de métodos de clusterização por modelagem (Fraley & Raftery, 1998).

Dessa forma, a Análise de Clusters por Modelagem (ACM) é o método que apresenta a maior precisão nos seus agrupamentos e estimativas, baseando-se em modelos de probabilidade e inferência estatística. Esse método de clusterização faz uso de distribuições normais centradas nas médias dos pontos – que representam as unidades de observação de cada cluster – e associa uma probabilidade de cada ponto ( $p$ ) em volta de cada média – que representa o conjunto de unidades de observação de cada cluster – estar presente em determinado cluster. A unidade de observação será alocada naquele cluster que tiver a média mais próxima de si (Bouveyron et al., 2019).

Com isso, também é possível estabelecer o grau de incerteza ( $g$ ) de uma unidade de observação pertencer a um determinado cluster, por meio do seguinte cálculo  $g = 1 - p$ . Os resultados do grau de incerteza podem ser apresentados em representação gráfica para facilitar análises e interpretações, utilizando-se, ainda, critério de corte para evidenciar somente as

<sup>8</sup> Essa referência ao formato esférico deve-se ao fato de que as distribuições de unidades de observação nesse formato são mais fáceis de serem agrupadas, enquanto as com formatos mais estranhos conferem maior complexidade no processo de clusterização.

unidades de observação que ultrapassem o limite estabelecido pelo critério. Para a presente pesquisa, foram evidenciadas graficamente somente as IES que tinham o grau de incerteza ( $g$ ) acima de 0,25 – 25% de incerteza de pertencimento ao cluster.

Esse foi o método escolhido para a pesquisa, por causa dos seguintes aspectos: i) a melhor precisão que esse método apresenta nos resultados; e ii) a utilização de probabilidade no processo de seus cálculos, o que se torna interessante para o estudo de unidades de observação que estão incertas em determinados grupos e a verificação de *outliers* para a formação dos clusters.

Assim como qualquer análise estatística, a ACM gera resultados tomando como base o conjunto de dados utilizado. Por isso, as semelhanças e as diferenças identificadas entre as unidades de observação – que se constituem como fundamentais para fins de delimitação e composição dos clusters – são estabelecidas somente para o conjunto de dados analisados.

Sendo assim, cada variável inserida ou retirada do processo de análise tem potencial de alterar a delimitação e a composição do conjunto de clusters. Ainda, a utilização de determinadas variáveis na ACM pode gerar viés nesse processo, caso essas variáveis reflitam uma característica comum com forte poder agregador das unidades de observação, em decorrência da pouca variabilidade que apresentem diante das demais variáveis utilizadas na aplicação desse método.

Esse foi o caso dos primeiros resultados obtidos em testes preliminares, em que a utilização de variáveis relativas à organização acadêmica das IES levou a uma delimitação e a uma composição de clusters que expressavam a mesma estrutura formal prevista na legislação vigente, com algumas nuances diferenciadoras. Com isso, poder-se-ia ter a percepção de que as IES com mesmas categorias acadêmicas oficiais tenderiam a apresentar aspectos de funcionamento similares, o que empiricamente não se confirmava.

Em decorrência desse resultado preliminar, foram retomadas as discussões sobre que variáveis deveriam, de fato, ser usadas na ACM, quando se concluiu que: i) conceitualmente, as variáveis de categorização oficial das IES não deveriam estar dentre aquelas selecionadas para expressar seus aspectos de funcionamento; e ii) estatisticamente, a pouca variabilidade de dados dessas variáveis, diante das demais, levavam a unidades de observação a serem agrupadas em torno da semelhança expressada pela categoria acadêmica das IES.

## Tratamento das variáveis para a ACM

O tratamento das variáveis para fins de utilização na ACM foi objeto de tópico anterior, pois parte era atinente tanto à AF quanto à ACM, sendo mais assertivo apresentá-los de forma concentrada, em que podem ser encontradas informações sobre como proceder e que opções metodológicas devem ser consideradas. Entretanto, cabe retomar essa temática sucintamente nesse ponto, em que é apresentada uma perspectiva operacional mais abrangente, passível de aplicação da ACM em qualquer estudo.

Assim sendo, com foco na ACM, usualmente, são considerados os seguintes tratamentos de dados, a serem aplicados quando da efetiva necessidade diante das características do conjunto de dados utilizados na análise:

- a) *seleção das variáveis*, atinentes ao recorte do fenômeno a ser analisado, com base na literatura, na experiência dos pesquisadores e nas características apresentadas pelas variáveis, com vistas aos pressupostos para a utilização da ACM;
- b) *verificação da qualidade dos dados*, com vistas à variabilidade de cada variável e à correlação entre as variáveis, considerando os impactos indesejados da ausência de variabilidade ou de correlação no processo de cálculo, que geram ruídos nas estimações ou inviabilizam utilização da ACM ou parte de seus modelos de estimação; e
- c) *redimensionamento das variáveis*, para mitigar problemas no cálculo computacional relativos ao volume de dados, evitar o *overfitting* em processos em que esteja sendo usada amostra para se fazerem inferências sobre a população ou viabilizar a interpretação dos clusters em estudos que abranjam muitas variáveis.

## Escolha do tipo de modelo na aplicação da ACM

Para delimitar os agrupamentos, a ACM conta com 14 modelos de estimação, baseados em matriz de covariância, que utiliza, conforme exposto no Quadro 1, três parâmetros, relativos a: i) *volume* – que define se os clusters vão ou não apresentar aproximadamente a mesma quantidade de unidades de observação; ii) *formato* – que define se os clusters teriam ou não a mesma variância; e iii) *orientação* – que define se os clusters são forçados a estarem em um ou

múltiplos eixos, podendo esses ser horizontal, vertical ou diagonal (Boehmke & Greenwell, 2020; Fraley, 1999).

#### Quadro 1

##### *Parametrização da matriz de covariância para aplicação na ACM*

| MODELO | VOLUME   | FORMATO  | ORIENTAÇÃO     | IDENTIFICADOR NO MCLUST |
|--------|----------|----------|----------------|-------------------------|
| 1      | Igual    | Igual    | Sem orientação | EII                     |
| 2      | Variável | Igual    | Sem orientação | VII                     |
| 3      | Igual    | Igual    | Eixos          | EEI                     |
| 4      | Variável | Igual    | Eixos          | VEI                     |
| 5      | Igual    | Variável | Eixos          | EVI                     |
| 6      | Variável | Variável | Eixos          | VVI                     |
| 7      | Igual    | Igual    | Igual          | EEE                     |
| 8      | Igual    | Variável | Igual          | EVE                     |
| 9      | Variável | Igual    | Igual          | VEE                     |
| 10     | Variável | Variável | Igual          | VVE                     |
| 11     | Igual    | Igual    | Variável       | EEV                     |
| 12     | Variável | Igual    | Variável       | VEV                     |
| 13     | Igual    | Variável | Variável       | EVV                     |
| 14     | Variável | Variável | Variável       | VVV                     |

Fonte: Boehmke e Greenwell (2020)

As escolhas pelo melhor modelo a ser utilizado na ACM e pela quantidade ótima de clusters podem, como critério estatístico, basear-se no Critério de Informação Bayesiano (BIC), em que o maior valor dessa medida de ajuste do modelo ao conjunto de dados indica a melhor opção a ser adotada (Scrucca et al., 2023).

Então, para utilizar o BIC, o primeiro passo seria: i) determinar o número máximo de clusters ( $m$ ), devendo ser o menor possível com vistas aos objetivos para a pesquisa e à praticidade das análises dos clusters, tomando-se como base, quando possível, algum referencial da literatura sobre fenômeno objeto da ACM; e ii) gerar as medidas de ajuste para os 14 tipos de modelos de estimação, no intervalo de 2 a  $m$  clusters (Fraley & Raftery, 1998).

Considerando o grau de subjetividade que envolve a definição do número de clusters máximo, pode ser viável testar alguns intervalos de tamanhos diferentes para uma melhor análise, pois podem ser obtidos modelos e quantidades ótimas de clusters diferentes à medida que o  $m$  é alterado para intervalos diferentes.

Com base nos valores de BIC obtidos para o intervalo de quantidades de clusters testados, passa-se à análise para identificar o melhor modelo de estimação a ser utilizado na ACM e a quantidade ótima de clusters. Nessa direção, o pacote “*mclust*” apresenta os resultados do BIC para os 14 modelos de estimação para cada quantidade de clusters testada em um gráfico de linhas, permitindo, para o conjunto de dados analisados: i) a identificação dos modelos que são aplicáveis na ACM; ii) a comparação entre os resultados do BIC para se identificar o modelo que melhor se ajusta aos dados; e iii) a quantidade ótima de clusters para a adoção na ACM. Nesse gráfico, cada linha se refere a um modelo de estimação e o eixo das abscissas ( $x$ ) estão as quantidades de clusters testadas, sendo o ponto de máximo aquele atinente ao melhor modelo e quantidade ótima de clusters.

Conforme explicitado, houve necessidade de revisões da lista de variáveis selecionadas para a pesquisa e dos pressupostos estatísticos que permeiam as etapas de aplicação dos métodos AM na presente análise. A última decorreu, sobretudo, dos resultados obtidos na ACM, envolvendo a seleção de modelo e a quantidade de cluster, tendo em vista a presença de resultados inesperados para um bom conjunto de dados, como: i) ausência de resultados para diversos modelos de estimação, que limitava as possibilidades dessa escolha para a aplicação da ACM; e ii) um comportamento sempre crescente no valor do BIC, fazendo com que, quanto maior fosse o  $m$  – número máximo de clusters a ser testado –, melhor essa medida ficava, o que inviabilizava a escolha da quantidade ótima de clusters por meio do BIC.

Investigando-se essa ocorrência, reforçou-se a importância da etapa de coleta e tratamento de dados, com um destaque, em relação a esse tipo de problema, para verificar a qualidade das variáveis e as distribuições que essas apresentam, tendo-se como referência os pressupostos e os pré-requisitos para aplicação das AM em determinado estudo.

Como o objetivo central deste texto é a explicitação da metodologia definida e utilizada para o estudo de tipologias institucionais focado no projeto de pesquisa em andamento no Laboratório de Pesquisa em Ensino Superior (Lapes), os resultados dos processos de análise, as discussões sobre a composição dos clusters e suas caracterizações em relação aos aspectos comuns das IES que os compõem, envolvendo o processo de análise e os respectivos resultados, serão objeto de publicações futuras.

## Considerações finais

Foram apresentados os aspectos metodológicos das análises quantitativas envolvidas na pesquisa, abordando-se aspectos teóricos da estatística e escolhas metodológicas, feitas com vistas aos pressupostos e aos pré-requisitos das AM e às características dos dados utilizados para aproximação dos aspectos de funcionamento das IES da ES brasileira. O processo de análises e os resultados obtidos, a serem apresentados futuramente, terão por foco: i) as dinâmicas da ES brasileira destacadas por meio das dimensões evidenciadas com a AFE e ii) a tipologia institucional da ES brasileira baseada nos aspectos de funcionamento de suas IES, evidenciada por meio dos clusters identificados com a aplicação da ACM.

No processo de delineamento metodológico e de aplicação das AM, foi possível reforçar perspectivas e hipóteses consideradas na pesquisa, além de sistematizar conhecimentos sobre os métodos de AM e sua aplicação em estudos sobre determinado fenômeno. Nessa direção, reforçou-se a importância de que a aplicação requer um conhecimento que extrapole a simples replicação de experiências similares bem-sucedidas. Isso porque cada conjunto de dados leva a resultados parciais, obtidos ao longo das AM, que podem ser muito diferentes, demandando tratamentos de dados e decisões, por parte do pesquisador, pautadas nos pressupostos e pré-requisitos dos métodos e outros conhecimentos estatísticos.

Também foi reforçada a necessidade do cuidado que se deve ter em relação às etapas de coleta e tratamento de dados, assim como as demais etapas de um estudo, pois o mínimo equívoco é capaz de invalidar todos os resultados obtidos com a aplicação dos métodos de AM, mesmo que essa aplicação seja conduzida corretamente. Ainda, ratificou-se a importância de a seleção de variáveis estar sustentada na literatura sobre o tema, no conhecimento dos pesquisadores sobre o fenômeno e nas características dos métodos estatísticos a serem utilizados posteriormente. Isso ao se considerar que uma variável inadequada para o estudo pode gerar ruídos nos processos de estimação dos métodos de AM e inviabilizar ou enviesar a geração de seus resultados.

Foi possível confirmar a perspectiva de que a ES dos demais países abarcados pela pesquisa tendem a apresentar tipologias institucionais, baseadas nos aspectos de funcionamento das IES, diferentes da que será definida para o Brasil. Isso decorre da própria forma como a educação superior é estruturada oficialmente nos demais países e dos dados existentes sobre o

funcionamento de suas IES. Entretanto, considerando as características da ACM, ainda será necessário verificar a possibilidade de sua aplicação ao contexto dos demais países.

Por essa razão, reafirmam-se algumas decisões relativas aos eixos de ação da pesquisa enfocando as comparações entre os países, sendo as análises estatísticas pertencente a um desses eixos. As comparações previstas, com vistas à estrutura formal dos países e às tipologias institucionais definidas a partir dos aspectos de funcionamento das IES, são as realizadas: i) entre os países, envolvendo as estruturas e aspectos de funcionamento da educação superior previstos no ordenamento legal; e ii) dentro de cada país, envolvendo as tipologias institucionais identificadas a partir dos dados empíricos e a estrutura e aspectos de funcionamento da educação superior previstos no ordenamento legal.

Por fim, foi possível compreender que outras tipologias institucionais podem ser obtidas a partir de seleções de variáveis, referenciadas em outras nuances da atuação das IES ou em seus diferentes recortes de resultados, atribuindo importância do delineamento metodológico aqui apresentado para fins de aplicação em outras pesquisas que enfoquem dinâmicas e resultados de determinada ES.

## Referências

- Alzahrani, A. R. R., Beh, E. J., & Stojanovska, E. (2021). Model-based clustering with *mclust* R package: Multivariate assessment of mathematics performance of students in Qatar. In *24th International Congress on Modelling and Simulation*, Sydney, NSW, Australia, 5 to 10 December 2021. [mssanz.org.au/modsim2021](https://mssanz.org.au/modsim2021).
- Alon, S. (2009). The evolution of class inequality in higher education: Competition, exclusion, and adaptation. *American Sociological Review*, 74(5), 731–755.
- Barbosa, M. L. de O., & Santos, C. T. (2011). A permeabilidade social das carreiras do ensino superior. *Cadernos CRH*, 24(63), 535–554.
- Betarelli Junior, A. A. , & Ferreira, S. de F. (2018). *Introdução à análise quantitativa e aos conjuntos Fuzzy (fsQCA)*. Enap.

- Boehmke, B., & Greenwell, B. M. (2020). *Hands-on machine learning with R*. CRC Press.
- Bouvyron, C., Celeux, G., Govaert, G., & Lancelot, G. (2019). *Model-based clustering and classification for data science: With applications in R*. Cambridge University Press.
- Chizzotti, A. (2003). *Pesquisa em ciências humanas e sociais* (7. ed.). Cortez.
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, 10(1), 7.
- Crespo, A. A. (2002). *Estatística fácil* (18. ed.). Saraiva.
- Creswell, J. W. (2010). *Research design: Qualitative, quantitative, and mixed methods approaches* (3. ed.). Sage Publications.
- Croxford, L., & Raffae, D. (2014). The iron law of hierarchy? Institutional differentiation in UK higher education. *Studies in Higher Education*, 40(9), 1625–1640.
- Fraley, C. (1999). Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20, 270–281.
- Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8), 578–588.
- Fumasoli, T., & Huisman, J. (2013). Strategic agency and system diversity: Conceptualizing institutional positioning in higher education. *Minerva*, 51(2), 155–169.
- Garson, G. D. (2022). *Factor analysis and dimension reduction in R: A social scientist's toolkit*. Routledge.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine.
- Hair, J. J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados* (6. ed.). Bookman.
- Hair, J. J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8. ed.). Cengage Learning.
- Hongyu, K. (2018). *Análise fatorial exploratória: Resumo teórico, aplicação e interpretação*. Engineering Science.

- Huisman, J., Lepori, B., Seeber, M., Frölich, N., & Scordato, L. (2015). Measuring institutional diversity across higher education systems. *Research Evaluation*, 24(4), 369–379.
- IBM. (2021). Exploratory factor analysis: Extraction. Retrieved September 16, 2024. <https://www.ibm.com/docs/en/spss-statistics/beta?topic=analysis-exploratory-factor-extraction>
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). (2019). *Manual para classificação dos cursos de graduação e sequenciais: CINE Brasil*. Brasília, DF: INEP.
- Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6. ed.). Upper Saddle River, NJ: Prentice Hall.
- Laros, J. (2012). O uso da análise fatorial: Algumas diretrizes para pesquisadores. In Lígia Pasquali (Ed.), *Análise fatorial para pesquisadores* (cap. 7, pp. 163–184). Labore Editorial.
- Lucas, S. R., & Moore, M. R. (2001). Tracking inequality: Stratification and mobility in American high schools. *American Journal of Sociology*, 107(2), 538-540.
- Luna, S. V. (1998). *Metodologia da pesquisa: Princípios e técnicas*. Educ.
- Manly, B. F. J., & Alberto, J. A. N. (2017). *Multivariate statistical methods: A primer* (4th ed.). CRC Press.
- Matos, A.S., & Rodrigues, E. C. (2019). *Análise fatorial*. Enap.
- Rodrigues, L. A. L. (2022). *A estratificação horizontal nos cursos imperiais: os concluintes de engenharia, direito e medicina entre 2009 e 2017*. [Tese de Doutorado, Universidade Federal do Rio de Janeiro].
- Schwartzman, S., Silva Filho, R. L., & Coelho, R. R. A. (2021). Por uma tipologia do ensino superior brasileiro: teste de conceito. *Estudos Avançados*, 35(101), 153-188.
- Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2023). *Model-based clustering, classification, and density estimation using mclust in R*. CRC Press.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591-611.
- Teixeira, P., Rocha, V., Biscaia, R., & Cardoso, M. F. (2013). Competition and diversification in public and private higher education. *Applied Economics*, 45(35), 4949-4958.

Triola, M. F. (2013). *Elementary statistics* (12. ed.). Pearson.

Van Vught, F. A. (2009). Diversity and differentiation in higher education. In F. Van Vught (Ed.), *Mapping the higher education landscape: Towards a European classification of higher education* (pp. 1–16). Springer.

Vieira, A. de H. P. (2021). *Estratificação no ensino superior e ingresso no mercado de trabalho no Brasil, 2009-2015*. [Tese de Doutorado, Universidade Federal do Rio de Janeiro].

Yin, R. K. (2016). *Qualitative research from start to finish* (2. ed.). The Guilford Press.

Zuccarelli, C., Vieira, A., Mendonça, L., & Blanco, F. (no prelo). A formação dos oligopólios da educação superior brasileira. In *Políticas de educación superior en América Latina: Expansión, diversificación y equidad*. CLACSO.

## Notas

### Preprint:

Previamente publicado como preprint em 28/03/2025  
<<https://doi.org/10.1590/SciELOPreprints.11543>>

### Dados da submissão:

Submetido à avaliação em 22 de maio de 2025; aprovado para publicação em 31 de julho de 2025.

### Autor correspondente:

**Santos, Renato Augusto dos** - SIG, Quadra 4, lote 327, Brasília - DF, Brasil, 70610-908.

### Contribuições de autoria:

**Dos Santos, Renato Augusto** - Conceituação (Liderança), Análise formal (Igual), Investigação (Igual), Metodologia (Liderança), Recursos (Igual), Supervisão (Liderança), Validação (Igual), Visualização (Igual), Escrita - rascunho original (Liderança), Escrita - revisão e edição (Igual).

**Raposo, Bruno Marão** - Conceituação (Suporte), Curadoria de dados (Igual), Análise formal (Igual), Investigação (Igual), Metodologia (Igual), Recursos (Igual), Software (Liderança), Validação (Igual), Visualização (Igual), Escrita - rascunho original (Suporte), Escrita - revisão e edição (Igual).

**Carrijo, Maria Clara Silva Cesar** - Conceituação (Suporte), Curadoria de dados (Igual), Análise formal (Igual), Investigação (Igual), Metodologia (Igual), Recursos (Igual), Software (Suporte), Validação (Igual), Visualização (Igual), Escrita - rascunho original (Suporte), Escrita - revisão e edição (Igual).

### **Apoio e financiamento:**

**Conselho Nacional de Desenvolvimento Científico e Tecnológico:** 420395/2022-9

**Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro:** E-26/200.863/2021

**Society for Research into Higher Education:** BARBOSA\_RA22287SRHE

### **Disponibilidade de dados:**

Os autores disponibilizam os dados da pesquisa sob solicitação.

### **Revisão textual:**

Normalização bibliográfica (APA 7ª Ed.), preparação e revisão textual em português: Vera Lúcia Fator Gouvêa Bonilha <verah.bonilha@gmail.com>

Versão e revisão em língua inglesa: Francisco López Toledo Corrêa <francisco.toledocorrea@gmail.com>

### **Editora responsável:**

Helena Sampaio <<https://orcid.org/0000-0002-1759-4875>>