Estado da publicação: Não informado pelo autor submissor

IDEB, AS UNIDADES FEDERATIVAS E O CENSO ESCOLAR: UMA ANÁLISE NÃO SUPERVISIONADA

Igor Alves, Rodrigo Jesus, João Romanelli, Carlos Silveira

https://doi.org/10.1590/SciELOPreprints.6097

Submetido em: 2023-05-16
Postado em: 2024-03-01 (versão 2)
(AAAA-MM-DD)

Justificativa da versão: Given some suggestions and criticisms received, corrections and adjustments were made to the text to make it more understandable.

ARTIGO

IDEB, AS UNIDADES FEDERATIVAS E O CENSO ESCOLAR: UMA ANÁLISE NÃO SUPERVISIONADA

IGOR MOREIRA ALVES[1]
ORCID: https://orcid.org/0000-0002-1309-2448
<igor.moreira@gmail.com>
RODRIGO MARCOS DE JESUS[2]
ORCID: https://orcid.org/0000-0002-2125-8794
<rodrigomarcosdejesus@yahoo.com.br>
JOÃO PAULO ROQUIM ROMANELLI[1]
ORCID: https://orcid.org/0000-0002-4280-7350
<joaoromanelli@unifei.edu.br>
CARLOS HENRIQUE DA SILVEIRA[1]
ORCID: https://orcid.org/0000-0001-6979-8469
<carlos.silveira@unifei.edu.br>

[1] Universidade Federal de Itajubá. Itabira, Minas Gerais (MG), Brasil.
[2] Universidade Federal do Mato Grosso. Cuiabá, Mato Grosso (MT), Brasil.

RESUMO: Este trabalho multidisciplinar propõe novos olhares e reflexões sobre os dados da multifacetada realidade educacional brasileira. Baseou-se numa criteriosa metodologia de análise exploratória não supervisionada, sem expectativa de resultados, de modo a dar chance ao novo de emergir espontaneamente a partir dos dados. Atenção especial recebeu a visualização de dados num contexto multidimensional, como forma de destacar sensorialmente os padrões mais imanentes, sem comprometimento do rigor estatístico. Logo, tal pesquisa teve como principal objetivo compor um entrelaçamento inovador de técnicas que oferecessem, no contexto de dados educacionais, mais oportunidades para perguntas e questionamentos que respostas. A base de dados consistiu no Índice de Desenvolvimento da Educação Básica (Ideb) e no Censo Escolar, num escopo fechado no ensino fundamental de 58920 escolas públicas de todo Brasil. No centro da metodologia esteve uma Cópula Gaussiana casada com uma Análise Fatorial recursiva, após cuidadoso tratamento de valores ausentes, valores extremos, dependências lineares e seleção de atributos. Como resultado, obteve-se uma pluralidade de padrões, tanto conhecidos da literatura, como inéditos. A exemplo do Ideb com peso em mais de uma dimensão, implicando em heterogeneidade de associações envolvendo este índice, os atributos de infraestrutura e as unidades da federação. Conclui-se que a metodologia foi confiável e robusta no desafio de compor novas perspectivas de análise sobre dados educacionais, viabilizando inclusive a ampliação do escopo em estudos futuros.

Palavras-chave: Indicadores educacionais, ensino fundamental, análise exploratória de dados, modelos não supervisionados

IDEB, THE FEDERATIVE UNITS AND THE SCHOOL CENSUS: AN UNSUPERVISED ANALYSIS

ABSTRACT: This multidisciplinary work proposes new perspectives and reflections on the data from the multifaceted reality of Brazilian education. It was based on a meticulous methodology of unsupervised exploratory analysis, without any expectation of results, in order to allow the emergence of new insights spontaneously from the data. Special attention was given to data visualization in a multidimensional context, as a way to sensorially highlight the most immanent patterns, without compromising statistical rigor. Therefore, the main objective of this research was to create an innovative intertwining of techniques that, in the context of educational data, would offer more opportunities for questions and inquiries than answers. The database consisted of the Basic Education Development Index (Ideb) and the School Census, within a scope limited to elementary education in 58,920 public schools across Brazil. At the heart of the methodology was a Gaussian Copula paired with a recursive Factor Analysis, after careful treatment of missing values, outliers, linear dependencies, and feature selection. As a result, a plurality of patterns was obtained, both known from the literature and unprecedented. Like Ideb having weight in more than one dimension, implying a heterogeneity of associations involving this index, the infrastructure attributes, and the federal units. In conclusion, the methodology has shown to be reliable and robust in the challenge of composing new analytical perspectives on educational data, enabling the expansion of the scope in future studies.

Keywords: Educational indicators, elementary education, exploratory data analysis, unsupervised models.

IDEB, LAS UNIDADES FEDERATIVAS Y EL CENSO ESCOLAR: UN ANÁLISIS NO SUPERVISADO

RESUMEN: Este trabajo multidisciplinario propone nuevas perspectivas y reflexiones sobre los datos de la realidad multifacética educativa brasileña. Se basó en una meticulosa metodología de análisis exploratorio no supervisado, sin expectativas de resultados, para permitir que lo nuevo emergiera espontáneamente a partir de los datos. Se prestó especial atención a la visualización de datos en un contexto multidimensional, como forma de resaltar sensorialmente los patrones más inminentes, sin comprometer el rigor estadístico. Por lo tanto, el objetivo principal de esta investigación fue componer un entrelazamiento innovador de técnicas que ofrecieran, en el contexto de los datos educativos, más oportunidades para preguntas y consultas que para respuestas. La base de datos consistió en el Índice de Desarrollo de la Educación Básica (Ideb) y el Censo Escolar, en un ámbito cerrado a la educación básica en 58.920 escuelas públicas de todo Brasil. En el centro de la metodología se encontraba una Cópula Gaussiana combinada con un Análisis Factorial recursivo, después de un tratamiento cuidadoso de los valores faltantes, los valores atípicos, las dependencias lineales y la selección de características. Como resultado, se obtuvo una pluralidad de patrones, tanto conocidos de la literatura como inéditos. Por ejemplo, el Ideb con peso en más de una dimensión, implicando una heterogeneidad de asociaciones que involucran este índice, los atributos de infraestructura y las unidades federativas. Se concluye que la metodología fue confiable y robusta en el desafío de componer nuevas perspectivas analíticas sobre los datos educativos, permitiendo incluso ampliar su alcance en futuros estudios.

Palabras clave: Indicadores educativos, educación primaria, análisis exploratorio de datos, modelos no supervisados.

INTRODUÇÃO

Vive-se em uma era que Jim Gray (2009) chamou de o 4o paradigma da exploração científica: pesquisa dado-intensiva, multidisciplinar, envolvendo volumes cada vez maiores de dados, processados por recursos computacionais em crescente sofisticação. Nesse contexto, como projetar modelos de análises capazes de extrair informação inédita e confiável a partir de bases de dados em ritmo acelerado de crescimento, mudança e complexidade? Principalmente, como pôr em relevo informações que induzam novos questionamentos, novos olhares sobre "o que se sabe que não se sabe" (*known unknowns*) ou mesmo tragam do recalque "o que não se sabe que não se sabe" (*unknown unknowns*) (Lakkaraju *et al.*, 2017, p. 1)? Os dados, per si, estão cada vez mais abundantes e acessíveis, mas como bem colocou Mark Abbott (2009, p. 113-14, tradução nossa): "não é mais a falta de dados que nos limita, mas a falta de insights." [1]

No Brasil, na área educacional, há uma profusa e sistemática produção de dados capitaneada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), dos quais destacam-se os gerados pelo Censo da Educação Básica (ou Censo Escolar) e pelo Sistema de Avaliação da Educação Básica (Saeb). Considerando que essas bases podem ser integradas a muitas outras fontes de dados igualmente intrincadas, volumosas e dinâmicas, envolvendo aspectos sociais, políticos e econômico-financeiros georreferenciados, tem-se uma excelente oportunidade para experimentar e avaliar técnicas de análise dado-intensivas na interface entre as ciências da educação e a ciência dos dados.

Essa integração ganha uma relevância especial no caso brasileiro, em função das desigualdades sociais e regionais, e seus reflexos condicionantes sobre os insumos e processos educacionais, em geral mais acentuadas aqui que em países desenvolvidos (Alves, 2020; Silva, 2022). Dependendo do contexto, apenas boas práticas no âmbito da gestão escolar interna nem sempre são suficientes para gerar impactos nos índices de qualidade educacional, sendo necessário também um arcabouço de políticas externas mais amplas e estáveis que garantam o mínimo de planejamento, continuidade, infraestrutura e recursos financeiros (Pontes, 2016). Logo, uma visão mais holística e sistêmica que leve em consideração a complexidade que norteia a heterogeneidade das escolas brasileiras pode ser fundamental para discriminar os acoplamentos mais sinérgicos entre as melhores práticas internas e políticas externas.

Com esses desafios em mente, neste trabalho empregou-se um conjunto de métodos e algoritmos cuidadosamente entrelaçados de modo a fazer emergir dos dados padrões visuais e numéricos estatisticamente confiáveis com o mínimo possível de viés ou expectativas de resultados. Ou seja, empreendeu-se uma análise exploratória de dados não supervisionada (James *et al.*, 2021). Modelos estatísticos supervisionados são aqueles em que se separa uma variável resposta ou dependente de um conjunto de variáveis preditoras ou independentes. Regressões são exemplos clássicos de método supervisionado, pois há de antemão a expectativa de que uma variável possa ser explicada como função das demais. Nos modelos não supervisionados, não há essa expectativa, pois não há separação de variáveis. Todas são tratadas indiferenciadamente, e se há alguma relação entre elas, essa relação deve apresentar-se de maneira espontânea ou não dirigida.

Como forma de aferir a metodologia não supervisionada aqui proposta, fez-se uma reavaliação das relações entre o Índice de Desenvolvimento da Educação Básica (Ideb) e os dados do Censo Escolar, mas limitada aos anos iniciais do ensino fundamental e à infraestrutura de escolas públicas. O escopo da análise foi propositalmente contido para que fosse possível elaborar comparações mais detalhadas entre os resultados gerados e os encontrados na literatura, sem perder de vista o conhecimento e a experiência prévia dos autores. Que a metodologia fosse capaz, enfim, de reproduzir "o que se sabe que se sabe" como uma forma empírica de validação.

---

[1] No original: *we are no longer data-limited but insight-limited.*

Neste trabalho, houve uma preocupação especial com a visualização dos dados, no intuito de instigar a imaginação, a reflexão e novos insights. Foram empregadas algumas inovações visuais e gráficas no contexto multidimensional, genericamente aqui chamadas de RnVis. Oferece uma série de visualizações de dados, algumas mais qualitativas, outras mais quantitativas, com múltiplas combinações ou possibilidades de análises. Como num caleidoscópio, espera-se que diferentes padrões possam emergir da análise conforme orientações do olhar. Como pontuou Graham Johnson:

> A ciência está se tornando mais visual, tanto na maneira como pesquisadores interagem com os dados - análise, interpretação e levantamento de hipóteses a partir da visualização dos resultados, dos modelos e dos arranjos gráficos contextuais - como também da maneira como eles comunicam esses resultados.[2] (Como citado em Rinaldi, 2012, p. 895, tradução nossa).

Logo, este artigo teve como principal objetivo promover novos olhares e reflexões sobre a multifacetada realidade educacional brasileira, através da composição de um entrelaçamento inovador de técnicas estatísticas e visuais, que pudessem estimular a curiosidade e a dúvida, induzindo mais o levantamento de perguntas e questionamentos que a oferta de respostas per si. Nesse sentido, em seu desenho experimental foi considerado fundamental que a metodologia fosse capaz não só de reproduzir resultados já bem delineados pela literatura, mas também oferecesse relações inéditas ou não tão destacadas em estudos prévios.

LITERATURA DE REFERÊNCIA

Foram destacados trabalhos recentes que serviram de referência para este estudo, especialmente artigos contendo revisões sobre infraestrutura e Ideb, além daqueles que fizeram algum tipo de uso de modelos estatísticos não supervisionados na área educacional.

Uma excelente revisão da literatura envolvendo a utilização de dados do Censo Escolar, do Saeb e outras, para caracterizar a infraestrutura e suas relações com índices de avaliação educacional pode ser encontrada em Alves; Xavier (2018). Os autores também construíram indicadores de infraestrutura de escolas do ensino fundamental com base em modelos da Teoria da Resposta ao Item (TRI) e Análise de Componentes Principais (PCA). Os indicadores possibilitaram uma abordagem multidimensional da infraestrutura educacional, cujos resultados reforçaram padrões já conhecidos na literatura, como as diferenças por vezes acentuadas: entre escolas privadas, federais, estaduais e municipais; entre urbanas e rurais; entre regiões, especialmente o contraste Sudeste/Sul e Norte/Nordeste; entre estabelecimentos que compartilham ensino fundamental e educação infantil, daqueles com ensino fundamental e médio; e até entre as que ofertam apenas ensino fundamental anos iniciais das que ofertam também anos finais.

Alves *et al.* (2019) oferece outra extensa revisão, em que ressalta a falta de consenso sobre mensurações e avaliações da infraestrutura escolar. Para mitigar essa questão, a revisão orientou a proposta de um modelo conceitual para avaliação da infraestrutura do ensino fundamental com base no Censo Escolar e no Saeb, sumarizada em 6 dimensões ou condições: da área (ex: localização); de atendimento (ex: modalidades e etapas); básicas (ex: serviços públicos, instalações mínimas); pedagógicas (ex: equipamentos, recursos pedagógicos); de bem-estar (ex: conforto e ambiente); para equidade (ex: acessibilidade, atendimento especializado). Fizeram então um mapeamento de um conjunto seleto de atributos do Censo e Saeb nessas dimensões, tendo como ano base 2015. Constataram que as dimensões envolvendo condições pedagógicas e básicas foram as mais presentes, em contraponto às de bem-estar e equidade.

Outra análise importante aparece em Alves (2020), com foco nos perfis das desigualdades educacionais de 2007 a 2017. Além de um estudo comparativo entre as diferentes evoluções ao longo

[2] No original: *Science is becoming more visual, both in the manner that researchers interact with their data—visual analysis of results, visual interpretation of results, and hypothesis generation from visual results, models, and contextual visual ensembles—and in the way they communicate.*

dos anos de gênero, raça/cor e nível socioeconômico em leitura e matemática, também descreveu efeitos das escolas sobre o desempenho médio dos alunos, com base em modelagem estatística de (Soares et al., 2016). Encontraram evidências de efeitos positivos no incremento do nível dos alunos, aspectos como: liderança administrativa, a gestão participativa, coesão da equipe pedagógica e equipamentos.

Cabe ainda destacar o extenso relatório sobre inclusão, equidade e desigualdades de estudantes de ensino fundamental público do Brasil, publicado em parceria pela *The United Nations Educational, Scientific and Cultural Organization* - Unesco e Núcleo de Pesquisa em Desigualdades Escolares (Nupede), da Universidade Federal de Minas Gerais - UFMG (Alves, et al., 2022). Além de vastas análises estatísticas descritivas de matrículas conforme grupos sociais e diversas varíaveis discriminantes (como sexo, cor/raça, nível sócioeconômico etc) nos anos de 2013, 2015 e 2017, apresentou também visualizações espaciais georreferenciadas como forma de construir uma verdadeira geografia das desigualdades regionais. Agregou ainda algumas análises exploratórias de correlação entre indicadores, grupos sociais e variáveis discriminantes. Entre os muitos resultados desse relatório, dado o contexto deste artigo, pode-se destacar a correlação positiva encontrada entre indicadores de melhor infraestrutura com maior taxa de aprovação.

Com respeito ao uso de metodologias não supervisionadas em dados educacionais, destaca-se a Análise Fatorial (FA) com base em dados do Projeto Geração Escolar 2005 - Estudo Longitudinal sobre Qualidade e Equidade no Ensino Fundamental - Projeto Geres (Faccenda *et al.*, 2011). Mas, nesse artigo, a FA foi usada apenas para redução dimensional, como uma etapa preliminar à aplicação de outras técnicas, visando relacionar práticas pedagógicas e condições escolares com eficácia e equidade escolar, que foi o objeto da tese em Dalben (2014). Dados do Projeto Geres também foram utilizados para avaliar um índice socioeconômico das escolas da educação básica modelado conforme TRI (Alves *et al.*, 2014). A FA foi a metodologia escolhida em Lopes (2022), em que se investigou as associações entre dados do Saeb e Ideb no ano base 2017, mas apenas para anos finais de escolas estaduais do estado do Tocantins. Foram encontradas associações do Ideb com boa infraestrutura (ex: sala de aula), internet de qualidade para professores e alunos, laboratórios de ciências e informática, índice socioeconômico médio ou alto. Já Ramos (2022), num artigo de escopo ainda mais limitado, empregou PCA para levantar padrões envolvendo atributos do Censo e Saeb dos anos finais do ensino fundamental de escolas municipais de Ribeirão Preto/SP.

CONSIDERAÇÕES METODOLÓGICAS

Base de dados

Foram utilizadas duas bases de dados: do Censo Escolar (Inep, 2019) e do Ideb de 2005 a 2019 (Inep, 2020). A escolha do ano de referência 2019 deve-se por ser o que antecede a pandemia de COVID-19, pois não fazia parte do escopo deste trabalho um estudo de sua influência. Outro motivo é que 2019 foi um ano de grandes alterações nos atributos dos microdados de infraestrutura do Censo Escolar em relação aos anos anteriores, em que foram adicionadas 103 variáveis, removidas 38, e alteradas outras 5, segundo levantamento próprio. Sobre o Ideb, fez-se uso do campo IDEB, compondo-se dois subíndices: a nota final IDEB 2019, como a nota mais próxima e não superior a 2019; a variação do IDEB, como a diferença entre a nota mais próxima e não superior a 2019 e a mais próxima e não inferior a 2005. As avaliações IDEB são bianuais em anos ímpares, mas por razões diversas nem todas as escolas têm notas para todos os anos ímpares entre 2005 e 2019. A regra acima permitia incluir o máximo de escolas, mesmo que algum ano ímpar estivesse sem nota. Com esses dois índices almejava-se uma análise exploratória de dados que levasse em consideração não somente o desempenho Ideb das escolas em 2019, mas também eventuais associações à variação desse índice no período 2005 a 2019. Foram consideradas apenas escolas públicas que estavam ativas em 2019 e que

tinham pelo menos uma nota IDEB referente aos anos iniciais do ensino fundamental. Nesta etapa, a base unificou-se numa matriz de dados com 58935 escolas e 158 atributos.

Imputação de valores ausentes

Valores ausentes (NAs ou *Not Available*) na base podem ser problemáticos, pois podem não só causar mal funcionamento em algumas rotinas estatísticas, como enviesar resultados (Azur *et al.*, 2011). A opção de trabalhar apenas com escolas com registros completos também pode induzir vieses, principalmente se a distribuição dos valores ausentes não for aleatória (Van Buuren, 2018). Além do risco de jogar informação valiosa fora. Por exemplo, na matriz aqui utilizada, algum valor ausente afeta cerca de um quarto das escolas (15828) e um terço dos atributos (51), mas não mais que 1% das células da matriz como um todo. Transformar indiscriminadamente todo valor ausente em zero também não é solução. O mais recomendado pela literatura é a imputação ou substituição dos valores ausentes, conforme alguma previsão estatisticamente fundamentada (Nguyen, 2022). Mesmo que as substituições não façam sentido para uma dada escola isoladamente, ao menos espera-se que essas trocas não comprometam a distribuição estatística geral.

Neste trabalho, optou-se pelo método de imputação por Cópula Gaussiana, que foi considerado igual ou superior ao estado da arte em publicação recente (Zhao, Udell, 2020). Estatisticamente, uma cópula é uma função capaz de mapear dependências ou associações entre atributos, tendo em vista distribuições marginais arbitrárias quaisquer (Durante, Sempi, 2010). Isso faz com que as cópulas sejam particularmente interessantes na construção de distribuições multivariadas a partir de dados com diferentes tipos de atributos, sejam contínuos ou categóricos (He *et al.*, 2021). Numa Cópula Gaussiana, mapeiam-se essas distribuições marginais numa distribuição Gaussiana conjunta multivariada. A imputação leva em conta essa distribuição conjunta para estimar os valores ausentes. Ela ainda oferece a vantagem de retornar uma matriz de correlação Sigma entre atributos que independerá do seu tipo e distribuição. Essa matriz Sigma foi aproveitada em outras etapas da metodologia, conforme explicitado adiante.

Controle de valores extremos

Valores extremos (*outliers*) também podem ter efeitos nocivos sobre as análises, seja por não terem sentido no contexto em consideração, seja por influenciar estimativas de parâmetros, deturpar modelos estatísticos e enviesar resultados (Aguinis *et al.*, 2013). Neste estudo, valores fora do escopo numérico definido pelo dicionário de dados do Censo Escolar foram marcados como ausentes e deixados para imputação. Para as demais situações, especialmente para atributos com prefixo QT, foi empregado o método de agrupamento DBSCAN (Schubert *et al.*, 2017), que mapeia a densidade dos pontos num espaço multidimensional. Pontos extremos tendem a ficar mais espalhados, em regiões de baixa densidade, longe de aglomerados. Foram identificadas 15 escolas que possuíam valores extremos suspeitos, por exemplo, mais de mil salas em QT_SALAS_UTILIZADAS_ DENTRO. Pela falta de confiabilidade gerada e por serem poucas, essas escolas foram eliminadas da base.

Seleção de atributos e dependências lineares

A base original do Censo Escolar vinha com 234 atributos. Durante o processo de limpeza e adequação da base ao foco do trabalho, foram eliminados:

- atributos não relacionados às escolas públicas municipais ou estaduais, como IN_MANT_ESCOLA_PRIVADA_EMP;
- atributos redundantes, como IN_TABLET_ALUNO, já representando por QT_TABLET_ALUNO;

- atributos com dependência linear de outros, como IN_ORGAO_NENHUM e os demais IN_ORGÃO;
- atributos muito específicos, como as derivações de IN_ACESSIBILIDADE;
- atributos com prefixo DT e CO (exceção para CO_ENTIDADE e CO_MUNICIPIO);
- atributos com comunalidade menor que 0,5 (será explicado mais adiante);

O atributo categórico nominal SG_UF foi binarizado, criando-se atributos dicotômicos (*dummies*) para cada unidade da federação, com nomes prefixados por "UF" (exemplo: UF_MG, para Minas Gerais). Para evitar questões de colinearidade, a unidade da federação com menor número de escolas (Roraima) ficou representada quando todos os demais atributos binários de prefixo UF são zeros.

Duas notas importantes: primeiro, que foi uma decisão estratégica trabalhar com os atributos em sua representação original, sem agrupamentos em indicadores ou modelos conceituais, de modo que a metodologia não supervisionada pudesse sugerir mais livremente as aglutinações dos atributos. Segundo, que foi a única etapa da metodologia em que houve uma interferência mais supervisionada, maior parte para evitar falhas, dificuldades de operacionalização ou resultados espúrios nas rotinas estatísticas utilizadas.

Nesta etapa, a matriz de dados alcançou 58920 escolas e 184 atributos, sendo 8 atributos identificadores e 176 atributos de dados.

Matriz de correlação

A metodologia aqui desenvolvida demandava uma matriz de correlação confiável. O problema é que a matriz de dados é mista, valendo-se de atributos categóricos binários, categóricos ordinais e atributos contínuos. É sabido que a clássica matriz de correlação Pearson não é adequada para variáveis categóricas, pois tende a subestimar seus valores (Holgado–Tello, *et al.*, 2010).

Uma alternativa é construir as correlações customizadas para cada par de atributos (Revelle, 2009a): contínuo x contínuo (Pearson); contínuo x categórico (polisserial); categórico x categórico (policórica). Tentou-se montar essa matriz de correlação mista, mas acusou-se a presença de autovalores negativos ao aplicar técnicas de redução dimensional, o que implicaria na possibilidade de variâncias negativas e desvios padrões imaginários, algo sem sentido no nível semântico dos dados. Uma matriz assim é chamada de não positiva definida, e há formas de tentar torná-la positiva definida, mas não sem algum grau de controvérsia (Lorenzo-Seva; Ferrando, 2021).

Ante tudo isso, optou-se por usar a matriz de correlação (Sigma) advinda da imputação por Cópula Gaussiana. Como defendem os autores dessa proposta (Zhao; Udell, 2020), essa matriz concebe correlações otimizadas entre atributos latentes marginais, que levam em consideração o tipo do atributo (contínuo ou categórico), numa metodologia unificada e padronizada para todos eles. Estratégia similar já foi utilizada em outros contextos de aprendizado de máquina, com resultados competitivos perante o estado da arte em diferentes bases e métricas (He, *et al.*, 2021).

Redução dimensional

O que se espera de uma análise exploratória de dados não supervisionada é a identificação e caracterização de padrões ou estruturas não aleatórias no espaço multidimensional dos dados. Em específico, verificar como as escolas e seus atributos se organizam e se associam num hiperespaço de 176 dimensões de dados. Mas tanto a mente humana quanto certos algoritmos de aprendizado de máquina têm dificuldades em detectar padrões em espaços multidimensionais extensos. Técnicas de redução dimensional permitem reduzir esse hiperespaço para poucas dimensões, com o mínimo de perda informacional (Cunningham, 2008).

Neste trabalho, optou-se pela técnica de FA (Williams, *et al.*, 2010). A FA encontrou nicho nas ciências sociais, nos estudos psicométricos e certas análises estatísticas educacionais (Henson; Roberts, 2006). É utilizada principalmente quando se trata de dados gerados por questionários, e o sistema Educacenso não deixa de ser um questionário (eletrônico). Tal como a técnica correlata de PCA, a FA também considera as variâncias, mas separando-as em pelo menos 4 tipos: gerais, de grupo, únicas e de erro, sendo as duas primeiras tratadas como variâncias comuns (comunalidades) e as duas últimas como variâncias exclusivas (singulares ou *uniqueness*) (Revelle, 2009b). O foco da FA está em isolar as variâncias exclusivas, priorizando as variâncias comuns ou compartilhadas.

Para este trabalho, foi utilizada a FA recursiva implementada por Storopoli (2019), que elimina o atributo com menor comunalidade a cada passo, até que sobrem apenas os atributos com comunalidades maiores que um ponto de corte (no caso, 0,494). Ou seja, ficarão apenas os atributos que compartilham mais de 50% (inclusive) da sua variância com outros atributos.

Costuma ser uma ênfase comum em estudos de FA evitar carga fatorial cruzada ou *crossloading* (Costello; Osborne, 2005), como consequência do princípio da estrutura mínima introduzido por Thurstone (1954). Em nome dessa parcimônia, praticam-se regras práticas como deixar apenas o maior valor de carga para cada atributo, zerando ou ignorando o resto (Revelle, 2009b) ou a remoção de atributos com carga maior que um certo valor (como 0.3) em dois ou mais fatores (Samuels, 2016). O resultado tende a ser um conjunto estrito de atributos únicos a cada fator, simplificando a interpretação. Mas, a que custo? Ao custo de subestimar a realidade em análise, no geral, intrinsecamente complexa (Ertel, 2011). Neste trabalho, optou-se por deixar a complexidade manifestar-se livremente na FA, não tratando as cargas fatoriais cruzadas, seguindo o defendido por Ertel (2013).

Cabe destacar, no entanto, que não foi adotada a rotação "varimin" proposta por Ertel, apenas foi permitida a carga cruzada na rotação "oblimin". A escolha dessa rotação se deu por admitir soluções tanto ortogonais quanto não ortogonais (oblíquas). E é apropriada quando suspeita-se que possam existir correlações entre os fatores (Watkins, 2018). É importante salientar que as rotações não interferem nos aspectos básicos resultantes da FA, como as variâncias acumuladas e as comunalidades. Visa apenas facilitar a visualização e interpretabilidade dos resultados (Costello; Osborne, 2005).

Toda a ciência dos dados deste trabalho foi desenvolvida em linguagem R. Detalhes das funções, parametrizações e pacotes utilizados encontram-se no Quadro 2 do Apêndice.

RESULTADOS E DISCUSSÕES

Preliminares da FA

A literatura recomenda pelo menos dois testes para verificar a adequação da base para a FA (Watkins, 2018): teste de Kaiser-Meyer-Olkin (KMO) e teste de Bartlett. O primeiro mede o grau de propensão dos atributos em ter variâncias compartilhadas uns com os outros. Um KMO próximo de um indica adequação para o uso da FA. O segundo testa se os valores da matriz de correlação são desvios significativos de uma matriz identidade, um pressuposto desejado. Obteve-se um KMO médio de 0,95, com um mínimo de 0,70 e máximo de 1,00. O teste de Bartlett deu p-valor de zero, sugerindo diferenciação ante uma matriz identidade. Ambos os testes indicam que a matriz de dados é fatorável.

Outro ponto importante seria quantos fatores extrair para uma FA, algo debatido e controverso na literatura (Van Der Eijk; Rose, 2015). A recomendação geral é contrastar diferentes métodos, deixando claro os critérios adotados na escolha. Aqui, foram utilizados três métodos (Revelle, 2009b): 1) método do cotovelo (*elbow*) ou do paredão (*scree*); 2) critério do Parcial Médio Mínimo (*Minimum Average Partial criterion* - MAP); 3) Análise Paralela. Uma descrição mais detalhada desses métodos está no Apêndice (Figura 9). Foram encontrados valores de 10, 12 e 28, respectivamente. Optou-se pelo número mais conservativo de 10 fatores, de modo a evitar superdimensionamentos e por terem sido suficientes para explicar cerca de 82% da variância dos dados.

Os testes de confiabilidade para a consistência interna dos fatores usando alfa de Cronbach (1951) foram satisfatórios, oscilando entre 0,68 e 0,96, conforme exibido no Quadro 3 do Apêndice.

Atributos eliminados pela FA

O Quadro 1 elenca 49 atributos que foram eliminados após a FA recursiva de Storopoli (2019), por apresentarem variâncias mais exclusivas, com comunalidades abaixo de 0,50. Logo, tenderam também a guardar baixas correlações com os demais atributos não eliminados da base. Houve uma redução de quase 30% no número de atributos, passando de 176 para 127.

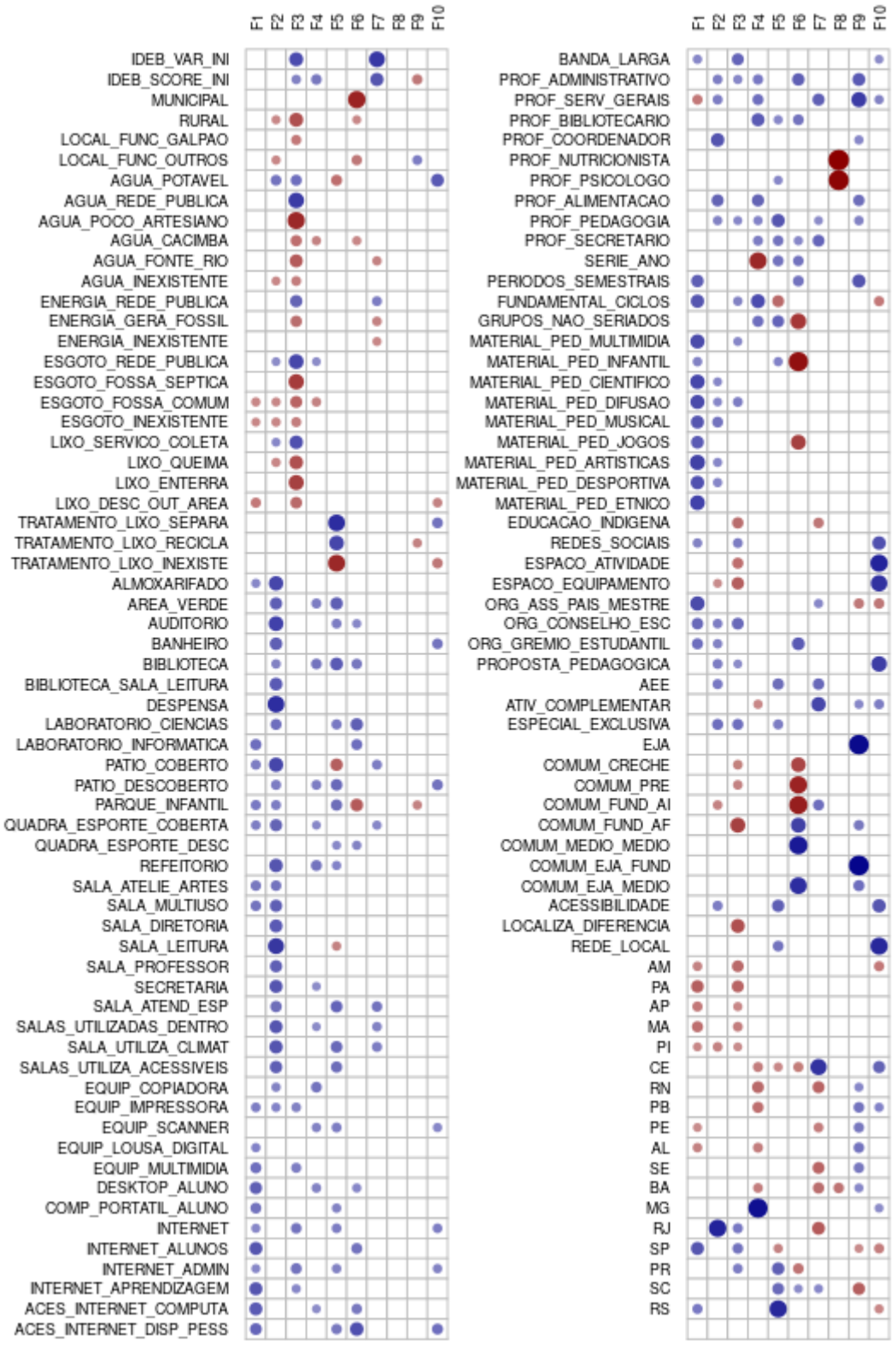
Quadro 1 - Listagem de atributos com comunalidades menores que 0,495.

ATRIBUTO	COM	ATRIBUTO	COM
RESERVA	0,001	AC	0,258
MEDIACAO_SEMIPRESENCIAL	0,009	ORGAO_ASS_PAIS	0,273
LOCAL_FUNC_SOCIOEDUCATIVO	0,023	PROF_SEGURANCA	0,273
ALIMENTACAO	0,029	PREDIO_COMPARTILHADO	0,287
FORMACAO_ALTERNANCIA	0,030	DEPENDENCIAS_OUTRAS	0,304
COMUM_EJA_PROF	0,035	RO	0,308
VIVEIRO	0,044	INTERNET_COMUNIDADE	0,309
PROF_SAUDE	0,052	ENERGIA_RENOVAVEL	0,317
LOCAL_FUNC_UNID_PRISIONAL	0,052	TABLET_ALUNO	0,324
REGULAR	0,067	GO	0,335
SALA_REPOUSO_ALUNO	0,068	MODULOS	0,336
LIXO_DESTINO_FINAL_PUBLICO	0,089	PROF_FONAUDIOLOGO	0,344
EXAME_SELECAO	0,114	LOCAL_FUNC_PREDIO_ESCOLAR	0,349
COMUM_PROF	0,131	SALA_MUSICA_CORAL	0,363
PISCINA	0,135	MATERIAL_PED_INDIGENA	0,379
TO	0,151	ORGAO_OUTROS	0,383
MT	0,130	TERREIRAO	0,389
SALA_ESTUDIO_DANCA	0,154	DORMITORIO	0,397
LOCAL_FUNC_SALA_OUTRA_ESC	0,174	TRATAMENTO_LIXO_REUTILIZA	0,414
DF	0,187	PROFISSIONALIZANTE	0,418
ES	0,190	EQUIP_PARABOLICA	0,422
COMUM_MEDIO_NORMAL	0,206	PROF_MONITORES	0,448
SALAS_UTILIZADAS_FORA	0,206	MATERIAL_PED_CAMPO	0,460
COMUM_MEDIO_INTEGRADO	0,227	COZINHA	0,476
MS	0,229		

Fonte: Elaborado pelos autores.
Nota: COM é abreviação de comunalidade. Prefixos removidos para melhor visualização.

Notam-se alguns padrões. Muitos dos atributos eliminados parecem especificidades, como: local de funcionamento socioeducativo (0,1% das escolas) ou tem piscina (0,6% das escolas). Outros, ao contrário, generalidades: se oferece alimentação (99,8% das escolas) ou funciona em prédio escolar (99,6% das escolas). Mas nem toda especificidade ou generalidade foi removida. Por exemplo: energia inexistente (0,3% das escolas) e presença de banheiro (98,8%) não apresentaram comunalidade baixa e não foram eliminados.

Figura 1 - RnVis geral com indicação visual dos coeficientes (carga) da matriz padrão resultante da FA.



Fonte: Elaborado pelos autores.
Notas: Azul indica carga positiva; vermelho, negativa.
Tamanho dos círculos proporcional ao valor da carga.

Correlações compõem uma medida invariante em escala. Guardam similaridade semântica e matemática com o ângulo entre dois vetores. E o valor do ângulo independe do tamanho ou da escala dos vetores. Logo, o que deve ter orientado a eliminação ou não de atributos por baixa comunalidade não foi tão somente efeitos de escala como especificidades e generalidades, mas talvez, principalmente, a presença de estruturas profundas nas correlações encontradas pela FA.

Todavia, o padrão mais curioso ocorreu com relação às unidades da federação com baixa comunalidade: todas da região Centro-Oeste: Mato Grosso (MT), Distrito Federal (DF), Mato Grosso do Sul (MS) e Goiás (GO); estados vizinhos a esta região: Tocantins (TO) e Roraima (RO), além de Espírito Santos (ES) e Acre (AC). Importante salientar que a eliminação dessas unidades não deve ser interpretada como entes federativos sem relevância, apenas sugere que não foram encontradas variâncias compartilhadas significativas com os atributos do Censo Escolar. Ou que não bastaram esses atributos para estabelecer um perfil de relações estatisticamente confiável para essas unidades. Mas por que justo essa região e demais estados destacados? Alves e Xavier (2018) relataram que a região Centro-Oeste ocupou posição intermediária em seus indicadores de infraestrutura, com Sudeste/Sul e Norte/Nordeste ocupando valores mais extremos. Estaria essa região tendo baixa comunalidade por estar mais próxima das medianas gerais dos dados? Eis uma questão em aberto que precisará ser melhor investigada em estudos futuros.

Resultados da FA

No RnVis da Figura 1, encontra-se representada a matriz de carga padrão (*pattern loading*), cujas células representam coeficientes (cargas) que marcam as relações entre atributos e fatores. Quanto maior o coeficiente em módulo, maior a área do círculo. Para efeito de melhor visualização, os coeficientes foram escalados pela raiz quadrada de seus valores absolutos, recuperando o sinal em seguida. E com intuito de despoluir a figura e destacar os coeficientes mais representativos, são exibidos somente os coeficientes com valor absoluto maior que 0.20. A literatura recomenda 0.3 (Samuels, 2016) como uma regra prática (*rule of thumb*), não como um limite teórico estabelecido. Ao longo deste trabalho essa regra foi relaxada, conforme o efeito ou adequação visual desejado. Se a projeção ocorrer no lado positivo do fator, o círculo estará destacado em azul; se sobre o lado negativo, em vermelho. Logo, atributos com mesma cor num mesmo fator guardam variâncias compartilhadas (associações). Com cores diferentes, podem estar em condições antípodas.

Alguns exemplos: atributos que continham o termo INTERNET ou MATERIAL_PED tenderam a ficar agrupados no mesmo fator F1 sob a cor azul; já os que continham o termo LIXO, mas sem serviço de coleta, como LIXO_QUEIMA e LIXO_ENTERRA, estão em vermelho no F2. Ainda com o termo LIXO, se são atributos que indicavam algum tipo de tratamento, como TRATAMENTO_LIXO_SEPARA, ficaram agrupados em azul no F3, mas se não há tratamento, como TRATAMENTO_LIXO_INEXISTENTE, em vermelho, sugerindo um antagonismo entre esses atributos. São relações que fazem sentido, são esperadas, e trazem confiabilidade aos resultados. O olhar também pode ser direcionado às linhas. Se assumirmos que cada fator agrega sentido pelos atributos que agrupa, a presença de um atributo em mais de fator pode ser interpretada como subconjuntos de escolas com características diferentes. O estado do Piauí (PI), por exemplo, marca presença nos três primeiro fatores, sugerindo subconjuntos de escolas piauienses agrupando atributos distintos.

Muitos outros padrões visuais emergem dessa Figura 1. Focando somente nos dois primeiros fatores (colunas F1 e F2), percebe-se que eles capturaram diferentes conjuntos de atributos, com poucas interseções. Isso é melhor percebido no RnVis ampliado da Figura 2, como se fosse um zoom da Figura 1, com informações também dos valores das cargas, apenas os maiores que 0,25 para que pudessem ser melhor visualizados na página.

Figura 2 - RnVis ampliado para os fatores F1 e F2.

ATRIBUTO	F1	ATRIBUTO	F2
LIXO_DESC_OUT_AREA	-0,26		
PROF_SERV_GERAIS	-0,26		
PA	-0,38		
AP	-0,26		
MA	-0,31		
LABORATORIO_INFORMATICA	0,31		
PATIO_COBERTO	0,25	PATIO_COBERTO	0,50
PARQUE_INFANTIL	0,26		
SALA_ATELIE_ARTES	0,27	SALA_ATELIE_ARTES	0,28
SALA_MULTIUSO	0,28	SALA_MULTIUSO	0,37
EQUIP_MULTIMIDIA	0,30		
DESKTOP_ALUNO	0,37		
COMP_PORTATIL_ALUNO	0,30		
INTERNET_ALUNOS	0,42		
INTERNET_APRENDIZAGEM	0,42		
ACES_INTERNET_COMPUTA	0,41		
ACES_INTERNET_DISP_PESS	0,35		
PERIODOS_SEMESTRAIS	0,38		
FUNDAMENTAL_CICLOS	0,44		
MATERIAL_PED_MULTIMIDIA	0,49		
MATERIAL_PED_CIENTIFICO	0,51		
MATERIAL_PED_DIFUSAO	0,48		
MATERIAL_PED_MUSICAL	0,42	MATERIAL_PED_MUSICAL	0,29
MATERIAL_PED_JOGOS	0,40		
MATERIAL_PED_ARTISTICAS	0,55		
MATERIAL_PED_DESPORTIVA	0,45		
MATERIAL_PED_ETNICO	0,53		
ORG_ASS_PAIS_MESTRE	0,50		
ORG_CONSELHO_ESC	0,33		
ORG_GREMIO_ESTUDANTIL	0,31		
SP	0,42		
RS	0,26		
		AGUA_POTAVEL	0,29
		ALMOXARIFADO	0,51
		AREA_VERDE	0,37
		AUDITORIO	0,53
		BANHEIRO	0,39
		BIBLIOTECA_SALA_LEITURA	0,40
		DESPENSA	0,66
		LABORATORIO_CIENCIAS	0,30
		PATIO_DESCOBERTO	0,26
		QUADRA_ESPORTE_COBERTA	0,37
		REFEITORIO	0,44
		SALA_DIRETORIA	0,41
		SALA_LEITURA	0,61
		SALA_PROFESSOR	0,37
		SECRETARIA	0,42
		SALA_ATEND_ESP	0,32
		SALAS_UTILIZADAS_DENTRO	0,42
		SALA_UTILIZA_CLIMAT	0,43
		SALAS_UTILIZA_ACESSIVEIS	0,38
		PROF_COORDENADOR	0,44
		PROF_ALIMENTACAO	0,36
		AEE	0,26
		ESPECIAL_EXCLUSIVA	0,31
		ACESSIBILIDADE	0,26
		RJ	0,72

Fonte: Elaborado pelos autores.
Notas: Azul indica carga positiva; vermelho, negativa. Cores escuras para cargas maiores que 0,30 em módulo. Cores claras, entre 0,25 e 0,30.

O fator F1 entrelaçou positivamente atributos mais relacionados à infraestrutura computacional e à várias modalidades de acesso e uso de Internet. E também: períodos semestrais, fundamental ciclos, vários tipos de material pedagógico, associação de pais e mestres, conselho escolar, grêmio estudantil. Tudo isso ficou mais amarrado aos estados de São Paulo (SP) e (em menor grau) ao Rio Grande do Sul (RS), em contraposição (em vermelho) a alguns estados do Norte e Nordeste.

Figura 3 - RnVis ampliado para os fatores F3, F4, F7 e F9.

ATRIBUTO	F3	ATRIBUTO	F4	ATRIBUTO	F7	ATRIBUTO	F9
RURAL	-0,47						
LOCAL_FUNC_GALPAO	-0,26						
AGUA_POCO_ARTESIANO	-0,70						
AGUA_CACIMBA	-0,31						
AGUA_FONTE_RIO	-0,41						
ENERGIA_GERA_FOSSIL	-0,32						
ESGOTO_FOSSA_SEPTICA	-0,57						
ESGOTO_FOSSA_COMUM	-0,36						
ESGOTO_INEXISTENTE	-0,25						
LIXO_QUEIMA	-0,46						
LIXO_ENTERRA	-0,55						
LIXO_DESC_OUT_AREA	-0,34						
EDUCACAO_INDIGENA	-0,32			EDUCACAO_INDIGENA	-0,28		
ESPACO_ATIVIDADE	-0,31						
ESPACO_EQUIPAMENTO	-0,38						
COMUM_FUND_AF	-0,54					COMUM_FUND_AF	0,27
LOCALIZA_DIFERENCIA	-0,47						
AM	-0,34						
PA	-0,36						
RJ	0,26			RJ	-0,41		
		SERIE_ANO	-0,70				
		CE	-0,25	CE	0,63		
		RN	-0,35	RN	-0,36		
		PB	-0,31			PB	0,29
		AL	-0,25			AL	0,28
		IDEB_SCORE_INI	0,28	IDEB_SCORE_INI	0,43	IDEB_SCORE_INI	-0,27
				SE	-0,35	SE	0,27
				BA	-0,32		
						ORG_ASS_PAIS_MESTRE	-0,27
						SC	-0,37
IDEB_VAR_INI	0,48			IDEB_VAR_INI	0,60		
AGUA_POTAVEL	0,31						
AGUA_REDE_PUBLICA	0,58						
ENERGIA_REDE_PUBLICA	0,37						
ESGOTO_REDE_PUBLICA	0,53						
LIXO_SERVICO_COLETA	0,45						
EQUIP_MULTIMIDIA	0,26						
INTERNET	0,28						
INTERNET_ADMIN	0,30						
BANDA_LARGA	0,36						
ORG_CONSELHO_ESC	0,35						
ESPECIAL_EXCLUSIVA	0,30						
SP	0,29						
		AREA_VERDE	0,25				
		BIBLIOTECA	0,28				
		PATIO_DESCOBERTO	0,25				
		REFEITORIO	0,30				
		EQUIP_COPIADORA	0,29				
		PROF_ADMINISTRATIVO	0,27			PROF_ADMINISTRATIVO	0,41
		PROF_SERV_GERAIS	0,30	PROF_SERV_GERAIS	0,37	PROF_SERV_GERAIS	0,55
		PROF_BIBLIOTECARIO	0,42				
		PROF_ALIMENTACAO	0,37			PROF_ALIMENTACAO	0,32
		PROF_SECRETARIO	0,25	PROF_SECRETARIO	0,35		
		FUNDAMENTAL_CICLOS	0,48				
		GRUPOS_NAO_SERIADOS	0,31				
		MG	0,86				
				PATIO_COBERTO	0,25		
				SALA_ATEND_ESP	0,28		
				SALA_UTILIZA_CLIMAT	0,26		
				AEE	0,33		
				ATIV_COMPLEMENTAR	0,50		
				COMUM_FUND_AI	0,30		
						PERIODOS_SEMESTRAIS	0,43
						EJA	0,90
						COMUM_EJA_FUND	0,93
						COMUM_EJA_MEDIO	0,31
						PE	0,27

Fonte: Elaborado pelos autores.
Notas: Azul indica carga positiva; vermelho, negativa. Cores escuras para cargas maiores que 0,30 em módulo. Cores claras, entre 0,25 e 0,30.

Já no fator F2 prevaleceu a infraestrutura construtiva ou arquitetônica, em termos de diversos tipos de salas, almoxarifado, auditório, banheiro, biblioteca, pátios, área verde etc, além de laboratório de ciências, turmas especiais exclusivas, sendo algo mais característico do estado do Rio de Janeiro (RJ). Nem F1 nem F2 mostraram associações relevantes com os índices do Ideb, indicando que, para um certo subconjunto de escolas, especialmente de SP, RS e RJ, as infraestruturas relatadas acima e esses índices foram mutuamente indiferentes.

Entretanto, a associação com o Ideb, tanto da taxa de variação entre 2015 e 2019 (IDEB_VAR_INI), quanto da nota de 2019 (IDEB_SCORE_INI), apareceu de forma diversa nos fatores F3, F4, F7 e F9. E apareceu espontaneamente, de forma não dirigida, em decorrência de tendências intrínsecas às correlações, como espera-se de modelos estatísticos não supervisionados. Os fatores F3 e F7 mostram-se ligados tanto à variação do Ideb quanto à nota do Ideb, enquanto os fatores F4 e F9 apenas à nota do Ideb. Interessante notar que o fator F9 indicou uma carga negativa em relação à nota do Ideb.

No RnVis ampliado da Figura 3, nota-se que o fator F3 foi marcado mais pela variação do Ideb, envolvendo atributos relacionados positivamente (azul): aos serviços públicos de água, energia, esgoto e lixo; internet com banda larga; classe exclusiva para alunos especiais; presença de conselho escolar; sobressaíram-se os estados de SP e RJ. No lado negativo de F3 (vermelho), praticamente o oposto: atributos que indicaram formas mais rudimentares do uso de água, energia, esgoto e lixo; localização diferenciada; educação indígena; espaços para atividades de integração escola-comunidade; rurais; presença dos estados do Amazonas (AM) e Pará (PA). Ou seja, F3 capturou os extremos da desigualdade brasileira, contrapondo o maior crescimento do Ideb de um Brasil tipicamente urbano, com estados representativos do Sudeste, daquele tipicamente rural com entes representativos do Norte, algo já bem conhecido da literatura (Alves; Xavier, 2018).

O fator F4 realçou um padrão diferente do F3. Para começar, ficou mais ligado à nota do Ideb em 2019 e não à sua variação. Positivamente, foi dominado com peso alto pelo estado de Minas Gerais (MG), em que se entrelaçaram os atributos que flexibilizam a seriação (grupos não seriados e fundamental ciclos); profissionais como bibliotecários e da alimentação; e alguma infraestrutura construtiva (refeitório, mas também área verde, biblioteca e pátio descoberto). No lado oposto, o atributo da seriação e alguns estados do nordeste. Do ponto de vista das escolas, pode-se supor a existência de um subgrupo de escolas tipicamente mineiras, com organização de um ensino não serial, que se alinham ao longo do eixo positivo de F4, com tendência de desempenho no Ideb 2019 acima da mediana.

O fator F7 também tem seu próprio padrão característico diferenciado. Do ponto de vista do Ideb, aparecem destacados tanto a variação quanto a nota do Ideb. A unidade da federação com vínculo forte aos dois índices é o Ceará (CE), em oposição ao RJ, Rio Grande do Norte (RN), Sergipe (SE) e Bahia (BA). Ademais, poucos atributos de infraestrutura se atrelam positivamente ao F7, com destaque para atividades complementares e Atendimento Educacional Especializado (AEE). Ainda apareceram: sala para AEE, pátio coberto e salas climatizadas. Não é surpresa que o estado do CE dominasse um fator, dado o seu notório destaque no cenário educacional brasileiro, em várias métricas (Alves, 2022). Mas, não deixa de ser um tanto surpreendente que, nesse fator em que há presença de ambos índices do Ideb, o CE tenha ficado atrelado a tão poucos atributos de infraestrutura.

Finalmente, o fator F9 teve a peculiaridade de apresentar uma nota Ideb negativa em associação com escolas que também abrigavam a modalidade Educação de Jovens e Adultos (EJA). Não por menos, apareceu em conjunção com o atributo de períodos semestrais. Em termos negativos, eram escolas que tendiam a não ter associação de pais e mestres. O F9 teve a tendência de atrelar-se a alguns estados do Nordeste, mas não Santa Catarina (SC).

Sobre os fatores ainda não comentados (F5, F6, F8, F10), nenhum deles fez associação ao Ideb, conforme pode ser visto no RnVis ampliado da Figura 4. F5 agrupou estados do Sul, em que destacam separação e reciclagem do lixo, atributos relacionados à acessibilidade e atendimento exclusivo, coerentes com padrões descritos no relatório da Unesco/UFMG (Alves, 2022). Com relação

à seriação, curiosamente houve associação positiva com grupos não seriados, mas negativa com fundamental ciclos.

O F6 representa no lado positivo principalmente atributos relacionados ao ensino médio, com predominância da esfera estadual (como esperado), mas não rural, com grêmio estudantil, períodos semestrais, laboratórios tanto de informática quanto ciências, algo percebido também por Alves; Xavier (2018). No lado negativo, material pedagógico infantil e de jogos, creches e/ou pré-escola, o que faz sentido num contexto fatorial em que o lado positivo domina o ensino médio em escolas estaduais. Pode parecer óbvio, mas é bom lembrar que esses padrões emergiram de forma automática dos dados, com o mínimo de intervenção humana, o que só atesta a confiabilidade das associações estatísticas encontradas pela metodologia aqui desenvolvida.

Figura 4 - RnVis ampliado para os fatores F5, F6, F8 e F10.

ATRIBUTO	F5	ATRIBUTO	F6	ATRIBUTO	F8	ATRIBUTO	F10
AGUA_POTAVEL	-0,31					AGUA_POTAVEL	0,41
TRATAMENTO_LIXO_INEXISTE	-0,69					TRATAMENTO_LIXO_INEXISTE	-0,28
PATIO_COBERTO	-0,38						
FUNDAMENTAL_CICLOS	-0,35					FUNDAMENTAL_CICLOS	-0,26
PARQUE_INFANTIL	0,32	PARQUE_INFANTIL	-0,40				
GRUPOS_NAO_SERIADOS	0,35	GRUPOS_NAO_SERIADOS	-0,60				
PR	0,39	PR	-0,29				
		MUNICIPAL	-0,72				
		LOCAL_FUNC_OUTROS	-0,27				
		MATERIAL_PED_INFANTIL	-0,87				
		MATERIAL_PED_JOGOS	-0,54				
		COMUM_CRECHE	-0,52				
		COMUM_PRE	-0,72				
		COMUM_FUND_AI	-0,77				
		CE	-0,26			CE	0,37
				PROF_NUTRICIONISTA	-0,98		
				PROF_PSICOLOGO	-0,93		
				BA	-0,28		
						ORG_ASS_PAIS_MESTRE	-0,28
						AM	-0,26
						SP	-0,25
TRATAMENTO_LIXO_SEPARA	0,67					TRATAMENTO_LIXO_SEPARA	0,29
TRATAMENTO_LIXO_RECICLA	0,52						
AREA_VERDE	0,38						
BIBLIOTECA	0,40	BIBLIOTECA	0,27				
LABORATORIO_CIENCIAS	0,26	LABORATORIO_CIENCIAS	0,38				
PATIO_DESCOBERTO	0,34					PATIO_DESCOBERTO	0,29
SALA_ATEND_ESP	0,35						
SALA_UTILIZA_CLIMAT	0,34						
SALAS_UTILIZA_ACESSIVEIS	0,32						
ACES_INTERNET_DISP_PESS	0,28	ACES_INTERNET_DISP_PESS	0,45			ACES_INTERNET_DISP_PESS	0,31
PROF_PEDAGOGIA	0,43						
PROF_SECRETARIO	0,28						
SERIE_ANO	0,30	SERIE_ANO	0,28				
AEE	0,32						
ESPECIAL_EXCLUSIVA	0,26						
ACESSIBILIDADE	0,38					ACESSIBILIDADE	0,44
REDE_LOCAL	0,29					REDE_LOCAL	0,71
SC	0,34						
RS	0,70						
		LABORATORIO_INFORMATICA	0,30				
		INTERNET_ALUNOS	0,30				
		ACES_INTERNET_COMPUTA	0,27				
		PROF_ADMINISTRATIVO	0,38				
		PROF_BIBLIOTECARIO	0,29				
		PERIODOS_SEMESTRAIS	0,30				
		ORG_GREMIO_ESTUDANTIL	0,40				
		COMUM_FUND_AF	0,54				
		COMUM_MEDIO_MEDIO	0,78				
		COMUM_EJA_MEDIO	0,67				
						BANHEIRO	0,28
						REDES_SOCIAIS	0,45
						ESPACO_ATIVIDADE	0,72
						ESPACO_EQUIPAMENTO	0,64
						PROPOSTA_PEDAGOGICA	0,58
						ATIV_COMPLEMENTAR	0,25

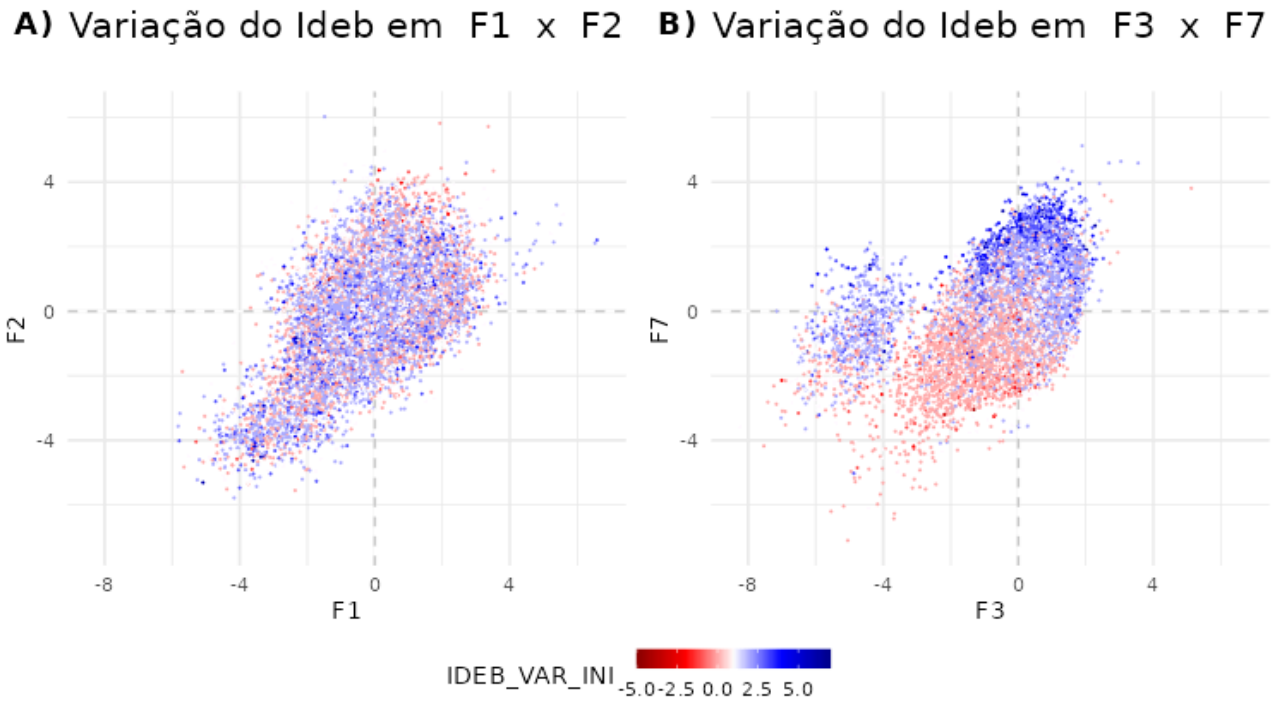
Fonte: Elaborado pelos autores.
Notas: Azul indica carga positiva; vermelho, negativa. Cores escuras para cargas maiores que 0,30 em módulo. Cores claras, entre 0,25 e 0,30.

O fator F8 teve a excepcionalidade de ter agregado com destaque apenas 3 atributos: com alto peso, nutricionistas e psicólogos; e em menor peso, a presença desses profissionais na BA. O fato desses dois atributos aparecerem isolados no lado negativo num único fator sugere que não guardam correlação relevante com nenhum outro atributo, a não ser entre eles mesmos. Também que as anticorrelações estão espalhadas nos demais atributos, não visíveis na figura, por estarem com cargas menores que o limite de 0,25.

Por fim, o fator F10 é marcado pelo compartilhamento de espaços internos para atividades comunitárias, uso de espaços e equipamentos externos em atividades de ensino e aprendizagem, uso de redes sociais, presença de rede local de dados com acesso sem fio à dispositivos pessoais, existência de proposta pedagógica, acessibilidade e em menor grau atividade complementar. Conta ainda com água potável, pátio descoberto, banheiro, e tende a separar lixo. Encontram-se menos presentes associações de pais e mestres, bem como fundamental em ciclos. Escolas com esse perfil são mais vistas no CE do que em SP e AM.

É intrigante que tenha aparecido novamente o CE, mas dessa vez sem a associação com o Ideb, como ocorreu com o F7. Pode-se interpretar que F10 e F7 representam subgrupos de escolas diferentes dentro do CE. É surpreendente que a FA tenha sido capaz de captar esses padrões diversos subjacentes e agrupá-los em dois fatores específicos. O subgrupo de escolas cearenses em F10 agregaram atributos que lhes são singulares, dissociando-as tanto de escolas com perfil do AM num extremo quanto de SP em outro. Mas, esse perfil de escolas do CE em F10 não sustenta correlação com o Ideb. Quem a faz, são as escolas do CE em F7. Por quê? Uma hipótese não testada é que talvez F10 contemple escolas cearenses que atingiram maior estabilidade nos índices do Ideb.

Figura 5 - Projeção das escolas nos fatores F1 x F2 e F3 x F7 conforme a variação do Ideb

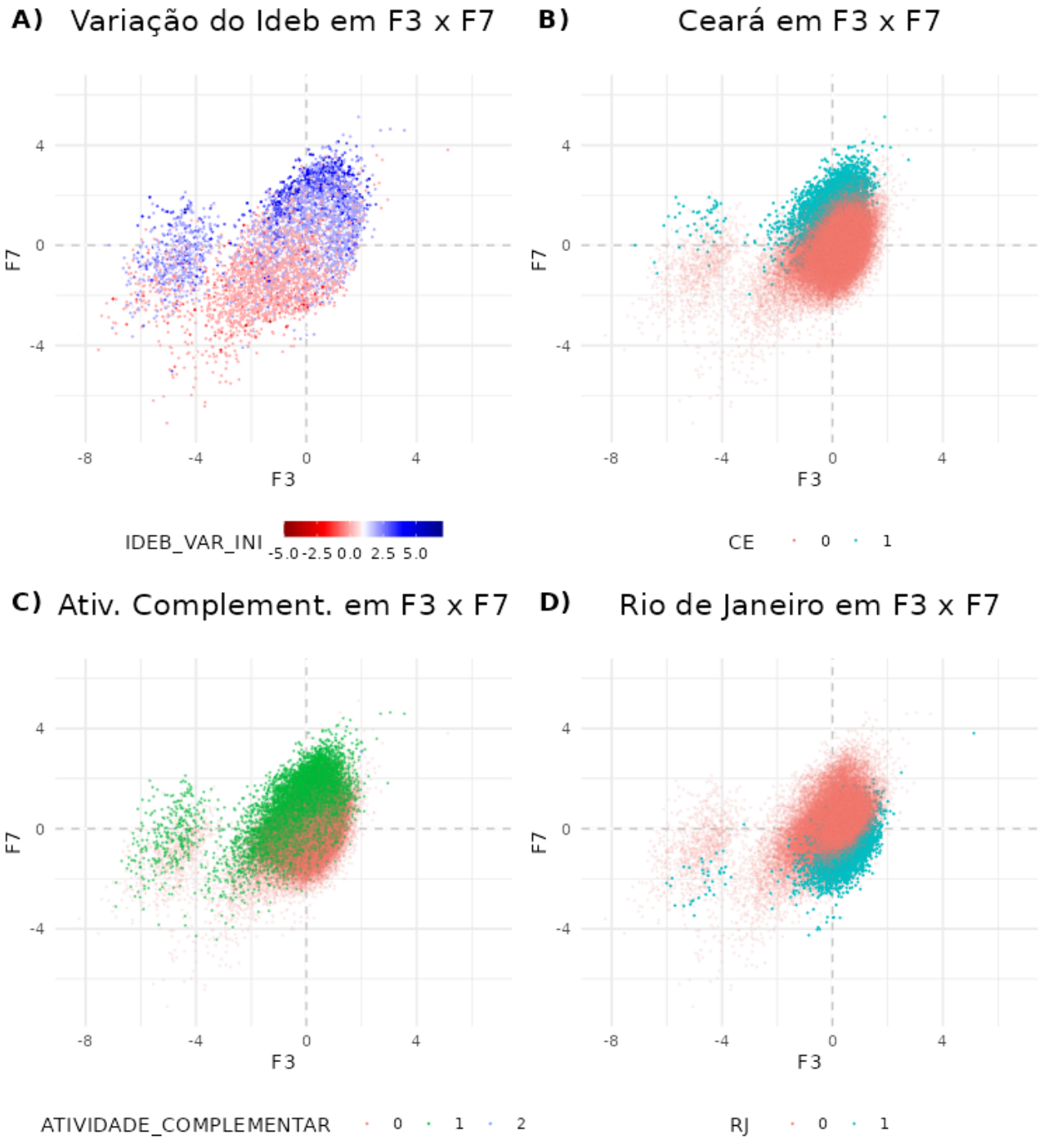


Fonte: Elaborado pelos autores.

Figura 6 - Projeção das escolas nos fatores F3 x F7 conforme atributos indicados nos gráficos

Até aqui pôde-se contemplar o hiperespaço dos atributos e fatores, realçando os padrões mais relevantes das associações entre atributos. Uma outra forma é olhar o hiperespaço entre

observáveis e fatores, ou seja, de como as quase 60 mil escolas ficam projetadas no espaço multidimensional dos fatores. Esse outro olhar permite dar destaque a novos padrões, além de complementar e expandir a compreensão dos já observados.

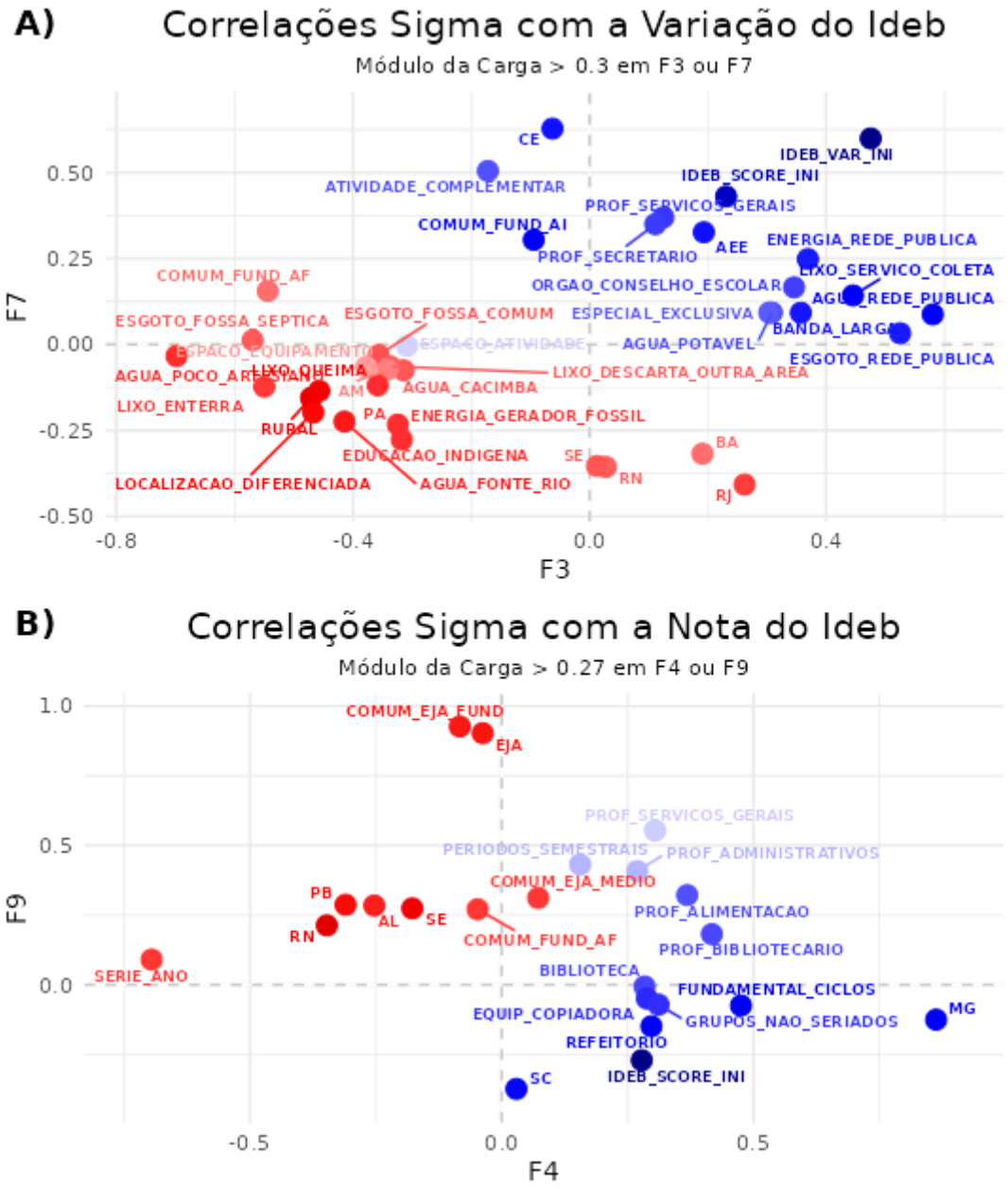


Fonte: Elaborado pelos autores.
Nota: Em atividade complementar, 0, 1 e 2 indicam: não oferece, não exclusivamente e exclusivamente.

A Figura 5 revela a distribuição das 58920 escolas no espaço dos fatores, considerando os pares F1 x F2 e F3 x F7. Cada ponto é uma escola, sendo as marcadas com tons de azul representam as com variação do Ideb acima da mediana (1.3), e com tons vermelho abaixo. Foi visto anteriormente que

os índices Ideb não se atrelaram significativamente aos fatores F1 e F2. De fato, na Figura 5-A, percebe-se uma distribuição de pontos azuis e vermelhos que parece aleatória, sem preferência por direções. O que já não é o caso dos pares F3 x F7 na Figura 5-B, pois há um nítido viés: quanto mais uma escola se posiciona nos valores mais altos dos eixos X e Y do gráfico, maiores as chances dela apresentar uma variação do Ideb acima da mediana.

Figura 7 - Projeção dos atributos nos fatores indicados conforme as correlações Sigma com Ideb.

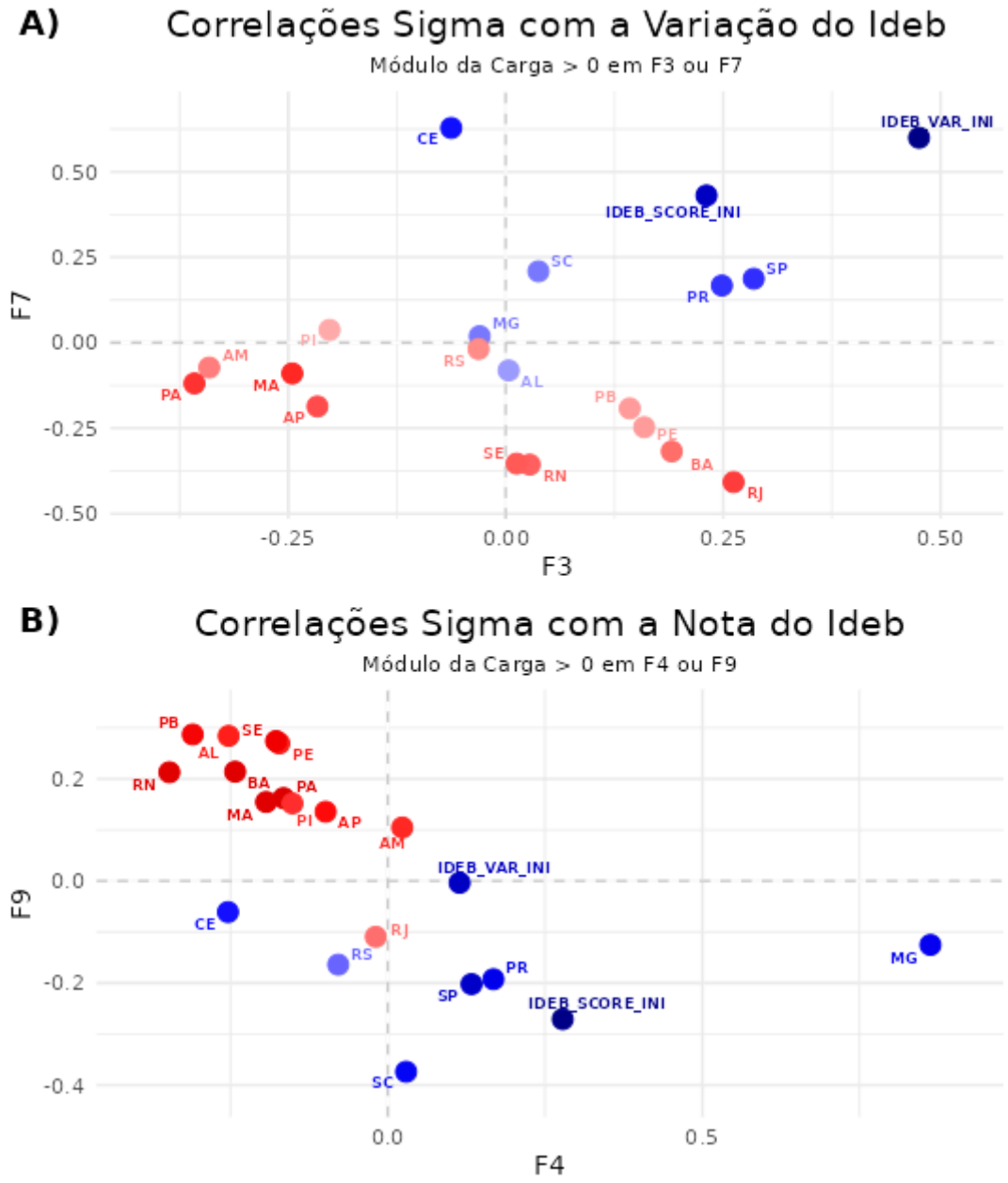


Fonte: Elaborado pelos autores.
Nota: Azul indica correlação Sigma positiva; vermelho, negativa.

Portanto, a assimetria de F3 e F7 na Figura 5-B sugere a possibilidade de três formas de escolas terem chances de possuir variação positiva do Ideb: apresentarem um F3 alto e um F7 baixo; apresentarem um F3 baixo, mas F7 alto; e apresentarem ambos F3 e F7 altos. E como cada fator tende a amarrar um conjunto de atributos diferentes, pode-se inferir a existência de diferentes padrões de

escolas com bom desempenho na variação do Ideb. Esse exemplo reforça o argumento de por que não se deve eliminar atributos com carga fatorial cruzada (carga em mais de um fator) em nome de uma estrutura mínima. Se isso fosse feito aqui, ficariam apenas os atributos carga F3 alta ou F7 alta, mas não ambas, como é o caso da variação do Ideb. Informação valiosa sobre o Ideb teria sido desprezada, em nome da simplificação.

Figura 8 - Projeção das unidades da federação nos fatores indicados conforme as correlações Sigma com Ideb.



Fonte: Elaborado pelos autores.
Nota: Azul indica correlação Sigma positiva; vermelho, negativa.

A Figura 6 vem auxiliar essa percepção. Olhando 6-A, 6-B e 6-C, notam-se, respectivamente: escolas com variação do Ideb acima da média, escolas do estado do CE e escolas com presença de atividade complementar, todas ocupando regiões mais positivas do fator F7 no eixo Y. Já

em 6-D, percebe-se que escolas do RJ tendem a ocupar uma região mais abaixo das escolas do CE, em parte no lado positivo de F3 e em parte no negativo de F7.

Os gráficos das Figuras 7 e 8 resumem as discussões levantadas até aqui, mas com um enfoque especial para a relação entre os fatores e as correlações da matriz Sigma dos atributos com os índices do Ideb, com as cores indicando o grau de correlação positiva (azuis) ou negativa (vermelhos). Vê-se que os pontos que estão mais próximos geometricamente dos atributos Ideb tendem a ter mais alta correlação Sigma com eles, e os mais distantes, baixa correlação, indicando que o hiperespaço dos fatores é de fato relacionado a um hiperespaço de correlações.

Em relação às unidades da federação, a Figura 8 realça alguns padrões já observados antes, mas provoca novas reflexões. Na Figura 8-A, fica claro como a variação do Ideb se apresenta em dois fatores, consequência de F3 e F7 altos. Os estados do Paraná (PR) e SP se alinharam mais juntamente ao F3, enquanto o CE ao F7. O estado de SC fica numa posição intermediária entre F3 e F7. Pode-se inferir que esses dados cresceram no Ideb a despeito de conjuntos de atributos de infraestrutura diversos.

Os estados do RJ, BA, Pernambuco (PE) e talvez Paraíba (PB) na Figura 8-A, em relação ao F3, seguiram padrões similares de PR e SP, mas diferentemente desses estados, apresentaram um F7 negativo, que os colocou em paridade com outros estados do Norte/Nordeste. Pode-se hipotetizar que tais estados contenham uma maior heterogeneidade no perfil de suas escolas, fragmentada em realidades diferentes. Os estados de MG, RS e Alagoas (AL) estão situados bem próximos da origem do plano cartesiano, indicando que foram indiferentes aos fatores F3 e F7.

Na Figura 8-B, o destaque fica para MG se isolando dos demais estados por um F4 bem elevado. Os estados do Norte/Nordeste ficaram mais agrupados no segundo quadrante. Os atributos variação do Ideb, AM e RJ estão mais próximos da origem, sugerindo maior indiferença em relação aos fatores F4 e F9.

Juntando F3, F7, F4 e F9 num hiperespaço de quatro dimensões, fica evidente que a nota do Ideb espalhou-se por essas quatro dimensões, ao contrário da variação do Ideb, que teve mais expressão em apenas duas dimensões.

CONCLUSÕES

O principal desafio deste trabalho foi conceber uma metodologia que fosse capaz de permitir novos olhares, novas reflexões e insights sobre os dados da multifacetada realidade educacional brasileira. Que gerasse mais oportunidades para perguntas e questionamentos que oferecesse respostas. A metodologia deveria ter como princípio uma posição cética ou neutra, com o mínimo de suposições cognitivas a priori. Os dados deveriam revelar seus padrões, como parte de um conjunto de observações sistemáticas e controladas sobre as características tabuladas das escolas, em consonância com os primeiros passos da metodologia científica clássica (Bacon, 1878). Este trabalho concentrou-se na mineração de padrões e não na formulação e teste de hipóteses com intuito de buscar explicações para tais padrões. Isso foi deixado para trabalhos subsequentes. A confiabilidade na metodologia seria aferida pela reprodução empírica de padrões já conhecidos da literatura. Essa confiança balizaria a ampliação do escopo de investigação também em estudos futuros.

Para tanto, desenhou-se uma criteriosa metodologia de análise exploratória de dados não supervisionada, sem expectativa de resultados, de forma a dar chance ao novo de emergir espontaneamente a partir dos dados. Estes foram devidamente pré-processados no intuito de imputar valores ausentes, controlar valores extremos, eliminar colinearidades e selecionar atributos relevantes. No centro dessa metodologia esteve a Cópula Gaussiana casada com uma FA recursiva. A cópula foi usada na imputação e na construção de uma matriz de correlação que operasse sobre tipos de atributos mistos, entre contínuos e categóricos. Essa matriz deu entrada à FA recursiva, utilizada tanto na eliminação de características pouco correlacionadas, quanto na evidenciação dos padrões de associação estatisticamente mais relevantes. Também foram utilizadas técnicas de visualização de dados num

contexto multidimensional que pudessem proporcionar maior evidência sensorial aos padrões mais imanentes. No conjunto, foram analisadas 58920 escolas públicas de todo o Brasil, caracterizadas segundo 176 atributos envolvendo infraestrutura e Ideb.

A decomposição FA revelou-se mais adequada em 10 dimensões, número suficiente para explicar 82% da variabilidade dos dados. Comunalidade abaixo de 0,50 afetou quase 30% dos atributos, que foram eliminados da base, dada sua baixa variância compartilhada. Entre os atributos eliminados estavam tanto aqueles mais específicos (ex: piscina) quanto gerais (ex: alimentação). Mas, nem todos atributos muito específicos ou gerais tiveram comunalidade baixa, indicando a presença de estruturas profundas nas correlações. Em relação às unidades da federação eliminadas, houve a surpresa de serem todas da região Centro-Oeste, além de vizinhos como TO e RO. Por que tão baixa comunalidade para essas unidades da federação, ainda não se tem resposta.

A metodologia foi capaz de reproduzir padrões já bem conhecidos da literatura. Evidenciou os extremos da desigualdade brasileira entre escolas públicas urbanas e rurais, entre as regiões Sul/Sudeste e Norte/Nordeste, tanto no perfil da infraestrutura, quanto no desempenho no Ideb, algo já destacado em Alves; Xavier (2018). No Nordeste, o estado do CE despontou com a sua reconhecida excepcionalidade regional, principalmente no que diz respeito aos avanços no Ideb. Boa parte dos padrões encontrados neste trabalho estão de acordo com relatório recente da Unesco/UFMG sobre o perfil dos estudantes das escolas públicas do ensino fundamental no Brasil (Alves, 2022). Nesse sentido, pode-se afirmar que a metodologia desenvolvida mostrou-se robusta ante os padrões já conhecidos.

Mas, com relação a outros resultados, teve-se dificuldades em rastreá-los na literatura. Embora alguns pareçam fazer sentido, dado o conhecimento e experiência prévias dos autores, outros causaram certa surpresa. Só para destacar alguns, entre os que parecem fazer sentido, estão as escolas que também operam a modalidade EJA com uma tendência de nota no Ideb menor que a mediana geral. Entre os não tão evidentes assim, está o estado de MG dominando solitariamente um fator (F4), em que se associaram também atributos que flexibilizam a seriação. Por que só MG com um destaque tão grande nesses atributos ante outros estados? Já o estado do CE, no fator (F7), destacou-se nos índices do Ideb, mas agregando poucos atributos de infraestrutura. Porém, o CE destacou-se também em outro fator (F10), este sem relação com Ideb, mas atraindo muito mais atributos de infraestrutura. O que exatamente isso significa? Além de curiosidades, como a exclusividade de profissionais nutricionistas e psicólogos controlando um fator (F8), e a tendência de estarem mais presentes na BA.

Cabe salientar que boa parte desses padrões só se tornaram evidentes pelo cuidado que se teve em respeitar a complexidade inerente aos dados, abrindo mão da busca por estruturas mínimas dentro da FA. Como argumenta Ertel (2011), a eliminação de atributos com carga fatorial cruzada é pura convenção em nome de um princípio de parcimônia. Mas isso não pode ser feito às custas de uma supersimplificação que viole estruturas naturalmente intrincadas. O mundo real raramente é simples.

É preciso que se olhe para os fatores como um espaço multidimensional, onde atributos e observáveis (escolas) podem se distribuir, se organizar, se projetar de forma complexa em várias dimensões, em várias direções. Por exemplo, mostrou-se aqui que a variação do Ideb tem peso em dois fatores, o que implica em pelo menos três subgrupos diferentes de atributos e escolas respondendo por essa variação. Que o CE está num fator dessa composição, e SP e PR em outra, tendo SC numa posição intermediária, mas todos apresentam-se atrelados a uma variação do Ideb positiva. Tais evidências parecem sugerir que o crescimento no Ideb ocorreu sob perfis de infraestrutura diferentes. Ou ainda que os atributos utilizados neste trabalho não foram suficientes para captar a eventual existência de um padrão comum subjacente ao desempenho no Ideb. Outro exemplo diz respeito a estados, como RJ, BA e PE, que têm em comum um fator com SP e PR, indicando partilharem certas características de infraestrutura. Mas se afastam deles por conta de outro fator, em que se posicionam mais alinhados a outros estados da região Norte/Nordeste. Parece que diferentes realidades fragmentaram os perfis das escolas públicas nesses estados com mais eloquência do que em outros. Se o princípio da estrutura mínima dentro da FA tivesse sido aplicado, talvez esses padrões nunca estariam sendo observados.

Há de se registrar algumas limitações deste trabalho. A primeira está no conjunto de dados utilizados, restritos aos atributos de infraestrutura do Censo Escolar e aos índices do Ideb, envolvendo anos iniciais do ensino fundamental de escolas públicas. Embora tenha sido uma limitação deliberada, ela tem seus impactos sobre as associações identificadas. A expansão do escopo e a integração com outras bases de dados pode reponderar os padrões e as evidências aqui destacados. Ressalta-se ainda que o grau de veracidade desses padrões está condicionado à qualidade dos dados advindos do Censo Escolar e do Ideb. Erros ou falhas na captação, armazenamento, anotação ou transformação desses dados na origem podem eventualmente ter enviesado todos os resultados apresentados.

Outra limitação é que este não foi um estudo longitudinal. Não fez parte do escopo um mapeamento da evolução dos atributos do Censo Escolar ao longo dos anos, de modo que concentrou-se atenção somente nas características das escolas em 2019, o ano que antecedeu a pandemia de COVID-19. Isso constitui um dos motivos pelos quais as bases de dados e metodologias empregadas, no momento, permitem apenas aferir correlações e associações, mas não causalidades.

Do ponto de vista da metodologia, embora ela tenha revelado ser capaz de reproduzir alguns padrões já conhecidos e apresentar outros novos nas bases utilizadas, há espaço para estudos mais aprofundados sobre alguns de seus aspectos. Por exemplo, a metodologia não aprofundou no estudo dos valores ausentes nem dos valores extremos nos dados do Censo Escolar. Também não esteve em foco uma análise comparativa mais detalhada entre as opções de matrizes de correlação disponíveis. Os padrões identificados são dependentes das correlações que deram entrada à FA, mas é relevante acentuar novamente que foi possível reproduzir resultados já bem conhecidos na literatura gerados por métodos diversos, alguns até óbvios, o que reforça a confiabilidade na matriz de correlação Sigma utilizada.

Por fim, este trabalho enfrentou o desafio de um diálogo multidisciplinar, conduzido por um filósofo da educação, um matemático e cientistas de dados. Foram muitas e diversas as realidades educacionais tocadas. Conforme já salientado, como um estudo exploratório, não fez parte do seu escopo formular e testar hipóteses para cada padrão observado, especialmente aqueles em que os autores tiveram dificuldades em rastrear na literatura, mesmo após diligente pesquisa. Cabe aos especialistas em cada área julgar se a metodologia e os padrões merecem destaques, ajustes ou refutações. Espera-se que os resultados apresentados possam dialogar com outras visões, estimulando o debate e o salutar exercício do contraditório.

REFERÊNCIAS

ABBOTT, Mark R. A New Path for Science? In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (org.). *The Fourth Paradigm*: Data-Intensive Scientific Discovery. Redmond, Washington: Microsoft Research, 2009, p. 111-116.
Disponível em:
<https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf>
Acesso em: 10/01/2024.

AGUINIS, Herman; GOTTFREDSON, Ryan K.; JOO, Harry. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*, v. 16, n. 2, p. 270–301, 2013. <https://doi.org/10.1177/1094428112470848>

ALVES, Maria Teresa Gonzaga; SOARES, José Francisco; XAVIER, Flavia Pereira. Índice Socioeconômico das Escolas de Educação Básica Brasileiras. *Ensaio*, v. 22, n. 84, p. 671–704, 2014. <https://doi.org/10.1590/S0104-40362014000300005>

ALVES, Maria Teresa Gonzaga; XAVIER, Flavia Pereira Indicadores Multidimensionais para Avaliação da Infraestrutura Escolar: o Ensino Fundamental. *Cadernos de Pesquisa*, v. 48, n. 169, p. 708–746, 2018. <https://doi.org/10.1590/198053145455>

ALVES, Maria Teresa; XAVIER, Flavia; PAULA, Túlio. Modelo Conceitual para Avaliação da Infraestrutura Escolar no Ensino Fundamental. *Revista Brasileira de Estudos Pedagógicos*, v. 100, n. 255, p. 297-300, 2019. <https://doi.org/10.24109/2176-6681.rbep.100i255.3866>

ALVES, Maria Teresa Gonzaga. Caracterização das Desigualdades Educacionais com Dados Públicos: Desafios para Conceituação e Operacionalização Empírica". *Lua Nova: Revista de Cultura e Política*, n. 110, p. 189–214, 2020. <https://doi.org/10.1590/0102-189214/110>

ALVES, M. T. G.; et al. *Inclusão, equidade e desigualdades entre estudantes das escolas públicas de ensino fundamental no Brasil*. Paris, França: Unesco, 2022.
Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000382175>. Acesso em: 10/01/2024.

AZUR, Melissa J. et al. Multiple Imputation by Chained Equations: what is it and how does it work? *International journal of methods in psychiatric research*, v. 20, n. 1, p. 40–49, 2011. <https://doi.org/10.1002/mpr.329>

BACON, Francis. (1878), *Novum organum*. Oxford: Clarendon press, 1878.
Disponível em: <https://books.google.com.br/books?id=tH4_AAAAYAAJ>. Acesso em: 10/01/2024.

COSTELLO, Anna B.; OSBORNE, Jason W. Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the most from your Analysis. *Practical Assessment, Research and Evaluation*, v. 10, n. 7, p. 1–10, 2005. <https://doi.org/10.7275/jyj1-4868>

CRONBACH, Lee J. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, v. 16, n. 3, p. 297–334, 1951. <https://doi.org/10.1007/BF02310555>

CUNNINGHAM, P. (2008), Dimension Reduction. In: CORD, M.; CUNNINGHAM, P. (org.). *Machine Learning Techniques for Multimedia*. Cognitive ed. Berlin, Heidelberg: Springer, 2008, p. 91–112. <https://doi.org/10.1007/978-3-540-75171-7_4>

DALBEN, Adilson. *Fatores Associados à Proficiência em Leitura e Matemática: uma Aplicação do Modelo Linear Hierárquico com Dados Longitudinais do Projeto GERES*. Tese (Doutorado em Educação). Campinas: Universidade Estadual de Campinas, 2014.
Disponível em: <https://repositorio.unicamp.br/Busca/Download?codigoArquivo=457001> Acesso em: 10/01/2024.

DURANTE, F.; SEMPI, C. Copula Theory: An Introduction. In: Jaworski, P., Durante, F., Härdle, W., Rychlik, T. (eds) *Copula Theory and Its Applications*. Lecture Notes in Statistics, vol 198. Berlin, Heidelberg: Springer, 2010, p. 3-31. <https://doi.org/10.1007/978-3-642-12465-5_1>

ERTEL, Suitbert. Exploratory Factor Analysis Revealing Complex Structure. *Personality and Individual Differences*, v. 50, n. 2, p. 196–200, 2011. <http://dx.doi.org/10.1016/j.paid.2010.09.026>

ERTEL, Suitbert *Factor Analysis*: Healing an Ailing Model. Göttingen: Universitätsverlag Göttingen, 2013. Disponível em: <http://www.varimin.com/downloads/ertel_factor.pdf> Acesso em: 10/01/2024.

FACCENDA, Odival; DALBEN, Adilson; FREITAS, Luiz. Capacidade Explicativa de Questionários de Contexto: Aspectos Metodológicos. *Revista Brasileira de Estudos Pedagógicos*, v. 92, n. 231, p. 246–267, 2011. <https://doi.org/10.24109/2176-6681.rbep.92i231.530>

GRAY, Jim. Jim Gray on eScience: A Transformed Scientific Method In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (org.). *The Fourth Paradigm*: Data-Intensive Scientific Discovery. Redmond, Washington: Microsoft Research, 2009, p. xvii–xxi.
<https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf>
Acesso em: 10/01/2024.

HE, Yi et al. Online Learning in Variable Feature Spaces with Mixed Data. In: IEEE International Conference on Data Mining, ICDM, v. 2021-Decem, p. 181–190, 2021. <https://doi.org/10.1109/ICDM51629.2021.00028>

HENSON, Robin K; ROBERTS, J Kyle. Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice. *Educational and Pschological Measurement,* v. 66, n. 3, p. 393–416, 2006. <https://doi.org/10.1177/0013164405282485>

HOLGADO–TELLO, Francisco Pablo et al. Polychoric versus Pearson Correlations in Exploratory and Confirmatory Factor Analysis of Ordinal Variables. *Quality and Quantity*, v. 44, n. 1, p. 153–166, 2010. <https://doi.org/10.1007/s11135-008-9190-y>

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). *Microdados do Censo Escolar da Educação Básica*, 2019.
Disponível em: <https://download.inep.gov.br/dados_abertos/microdados_censo_escolar_2019.zip>
Acesso em: 10/01/2024.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). *Ensino Fundamental Regular - Anos Iniciais (Atualizado em 18/09/2020)*, 2020.
Disponível em:
<https://download.inep.gov.br/educacao_basica/portal_ideb/planilhas_para_download/2019/divulgacao_anos_iniciais_escolas_2019.zip>. Acesso em: 10/01/2021.

JAMES, Gareth et al. *An Introduction to Statistical Learning*: with Applications: in R. 2nd. ed. New York: Springer, 2021. Disponível em: <https://www.statlearning.com/> Acesso em: 10/01/2024.

LAKKARAJU, Himabindu et al. (2017), Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration. In: 31st AAAI Conference on Artificial Intelligence, AAAI 2017, v. 31, n. 1, p. 2124–2132. <https://doi.org/10.1609/aaai.v31i1.10821>

LOPES, S. M. M. C. *Análise Fatorial Multivariada Aplicada na Avaliação Educacional das Escolas Estaduais de Ensino Fundamental do Estado do Tocantins*. Dissertação (Mestrado em Educação). Palmas: Universidade Federal do Tocantins, 2022. Disponível em: <https://repositorio.uft.edu.br/handle/11612/4081>. Acesso em: 10/01/2024.

LORENZO-SEVA, Urbano; FERRANDO, Pere J. Not Positive Definite Correlation Matrices in Exploratory Item Factor Analysis: Causes, Consequences and a Proposed Solution. *Structural Equation Modeling*, v. 28, n. 1, p. 138–147, 2021. <https://doi.org/10.1080/10705511.2020.1735393>

NGUYEN, Mike. A Guide on Data Analysis. Bookdown R package. E-book, 2022. Disponível em: <https://bookdown.org/mike/data_analysis/>. Acesso em: 10/01/2024.

PONTES, T. C. A. de. *The Politics of Education Reform in Brazilian Municipalities*. Thesis (Bachelor of Arts, Department of Government). Cambridge, USA: Harvard College, 2016.
Disponível em: <https://undergrad.gov.harvard.edu/files/undergradgov/files/comp_thesis_7.pdf>. Acesso em: 10/01/2024.

RAMOS, Mozart Neves et al. Uma Análise Estatística Multivariada do Desempenho das Escolas Municipais de Ribeirão Preto. *Ensaio: Avaliação e Políticas Públicas em Educação*, v. 29, n. 113, p. 857–873, 2021. <https://doi.org/10.1590/S0104-40362021002903286>

REVELLE, W. Covariance, Regression, and Correlation. In: *An Introduction to Psychometric Theory with Applications in R*. Personality Project. E-Book, 2009a.
Disponível em: <https://personality-project.org/r/book/>. Acesso em: 10/01/2024.

REVELLE, W. Constructs, Components, and Factor Models. In: *An Introduction to Psychometric Theory with Applications in R*. Personality Project. E-Book, 2009b.
Disponível em: <https://personality-project.org/r/book/>. Acesso em: 10/01/2024.

RINALDI, Andrea. More than meets the eye. Modern experimental techniques require increasingly sophisticated approaches to data visualization. *EMBO Reports*,, v. 13, n. 10, p. 895–899, 2012. <http://dx.doi.org/10.1038/embor.2012.135>

SAMUELS, Peter. *Advice on Exploratory Factor Analysis. Centre for Academic Success*. Report. Birmingham, England: Birmingham City University, 2016.
Disponível em: <https://www.open-access.bcu.ac.uk/6076/>. Acesso em: 10/01/2024.

SCHUBERT, Erich et al. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems*, v. 42, n. 3, p. 19:1-19:21, 2017. <https://doi.org/10.1145/3068335>

SILVA, Flávia Fernanda Santos. A Infraestrutura como Fator Associado ao Desempenho dos Alunos no Ensino Fundamental. *Revista Amazônida*: Revista do Programa de Pós-Graduação em Educação da Universidade Federal do Amazonas, v. 7, n. 1, p. 1–22, 2022. <https://doi.org/10.29280/rappge.v7i01.10607>

SOARES, José Francisco; ALVES, Maria Teresa Gonzaga; XAVIER, Flavia Pereira. Effects of Brazilian Schools on Student Learning. *Assessment in Education*: Principles, Policy and Practice. v. 23, n. 1, p. 75–97, 2016. <http://dx.doi.org/10.1080/0969594X.2015.1043856>

STOROPOLI, J. How to use Factor Assumptions. Github, 2019. Disponível em: <https://github.com/storopoli/FactorAssumptions>. Acesso em: 10/01/2024.

THURSTONE, L L. (1954), An Analytical Method for Simple Structure. *Psychometrika*. v. 19, n. 3, p. 173–182, 1954. Disponível em: <https://link.springer.com/content/pdf/10.1007/BF02289182.pdf>. Acesso em: 10/01/2024.

VAN BUUREN, Stef. *Flexible Imputation of Missing Data*. 2nd. ed., Chapman & Hall/CRC. E-book, 2018. Disponível em: <https://stefvanbuuren.name/fimd/>. Acesso em: 10/01/2024.

VAN DER EIJK, Cees; ROSE, Jonathan. Risky Business: Factor Analysis of Survey Data - Assessing the Probability of Incorrect Dimensionalisation. *PLoS ONE*, v. 10, n. 3, p. 1–31, 2015. <https://doi.org/10.1371/journal.pone.0118900>

WATKINS, Marley W. Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology*, v. 44, n. 3, p. 219–246, 2018. <https://doi.org/10.1177/0095798418771807>

WILLIAMS, Brett; ONSMAN, Andrys; BROWN, Ted. Exploratory Factor Analysis: A Five-Step Guide for Novices. *Journal of Emergency Primary Health Care*, v. 8, n. 3, p. 1–13, 2010. <https://doi.org/10.33151/ajp.8.3.93>

ZHAO, Yuxuan; UDELL, Madeleine. Missing Value Imputation for Mixed Data via Gaussian Copula. In: , 2020. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 636–646, 2020.
Disponível em: <https://dl.acm.org/doi/pdf/10.1145/3394486.3403106>. Acesso em: 10/01/2024.

CONTRIBUIÇÃO DOS AUTORES

Autor 1 - Administração do projeto, curadoria de dados, escrita - primeira versão, investigação, software e visualização.
Autor 2 - Análise formal, conceituação, curadoria de dados, escrita - revisão, supervisão, validação.
Autor 3 - Análise formal, conceituação, curadoria de dados, escrita - revisão, metodologia, supervisão, validação.
Autor 4 - Administração do projeto, análise formal, curadoria de dados, escrita – revisão, edição, investigação, metodologia, obtenção de financiamento, recursos, software, supervisão, visualização.

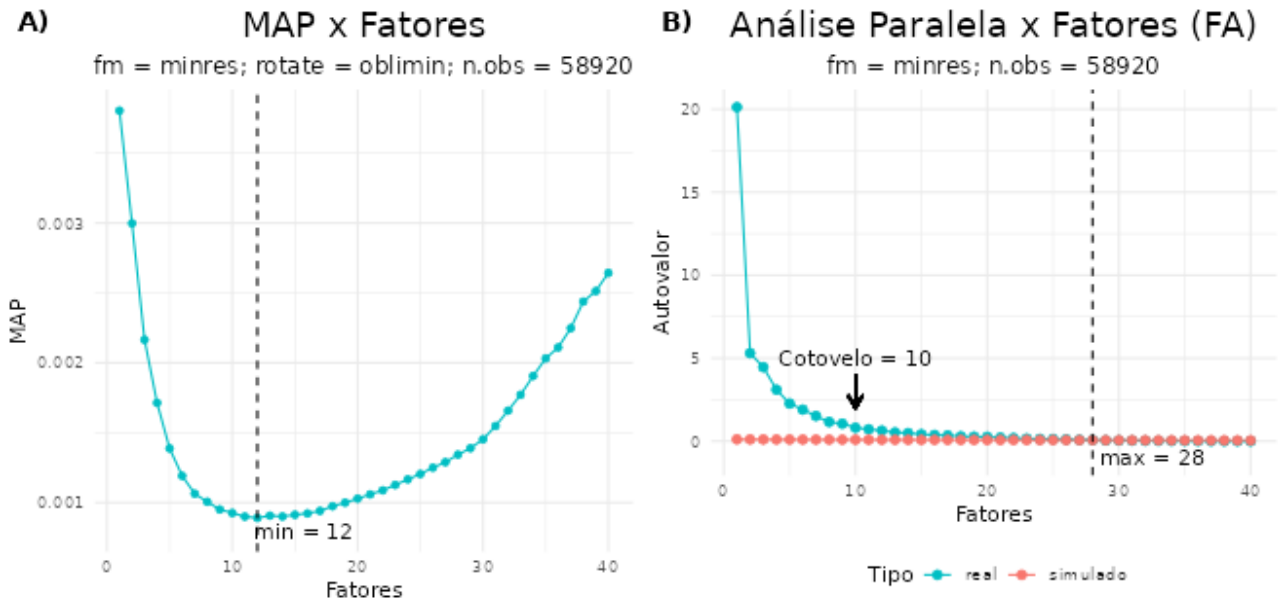
DECLARAÇÃO DE CONFLITO DE INTERESSE

Os autores declaram que não há conflito de interesse com o presente artigo.

APÊNDICE

Figura 9 mostra gráficos da estimação do número de fatores a serem usados na FA. Em Figura 9-A, tem-se a variação do critério MAP em função do número de fatores. Observa-se um ponto de mínimo em torno do número 12. Em figura 9-B, a variação dos autovetores com dados reais (azul) e dados de uma simulação aleatória (vermelho). O número 28 representa o número máximo de fatores em que os autovalores reais são maiores que os simulados randomicamente. A seta indica um possível cotovelo em torno de 10 fatores.

Figura 9 - Determinação do número de fatores por diferentes métodos.



Fonte: Elaborado pelos autores.

No Quadro 2, as variáveis "matrix" e "corr.matrix" correspondem às matrizes de dados e de correlação, respectivamente; e "keys" representam os atributos projetados nos fatores. Desenvolvido em: R version 4.2.2 Patched (2022-11-10 r83330), Rstudio 2022.07.2+576, Ubuntu 20.04.5 LTS.

Quadro 2 - Principais funções e pacotes do R e respectivos parâmetros utilizados.

TÉCNICA	FUNÇÃO/PARÂMETROS	PACOTE
Cópula Gaussiana	impute_GC(matrix, nlevel = 5, verbose = TRUE)	gcimputeR 0.1.1
Valores Extremos	dbscan(scale(matrix), eps = 20, minPts = 44)	dbscan 1.1-10
KMO	KMO(corr.matrix)	psych 2.2.5
Teste de Bartlett	cortest.bartlett (corr.matrix, n = 58920)	psych 2.2.5
MAP	VSS(corr.matrix, n = 40, fm = "minres", rotate = "oblimin", n.obs = 58920)	psych 2.2.5
Análise Paralela	fa.parallel(corr.matrix, n = 40, fm = "minres", n.obs = 58920)	psych 2.2.5

Análise Fatorial Recursiva	com_opt_sol(corr.matrix, nfactors = 10, rotate = “oblimin”, cutoff = 0.494, fm = “minres”, n.obs = 58920)	Adaptado de Storopoli (2019)
Projeção dos observáveis	factor.scores(matrix, fa$results, method = "tenBerge")	psych 2.2.5
alpha de Cronbach	alpha(corr.matrix, keys = keys, n.obs = 58920, n.iter = 1000)	psych 2.2.5

Fonte: Elaborado pelos autores.

Quadro 3 mostra o retorno de teste de confiabilidade alfa de Cronbach, com intervalo de confiança a 0,95 determinado por bootstrap com 1000 iterações. Foi calculado um alfa separado para cada fator.

Quadro 3 - Resultado do teste de confiabilidade alfa de Cronbach.

INT	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**
Inferior	0,947	0,949	0,955	0,864	0,888	0,908	0,760	0,534	0,692	0,726
Alfa	0,956	0,959	0,962	0,888	0,908	0,925	0,798	0,677	0,775	0,787
Superior	0,962	0,968	0,968	0,910	0,924	0,940	0,834	0,756	0,824	0,827

Fonte: Elaborado pelos autores.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.