

Uma Biblioteca em R para Análise de Dados Educacionais

Pedro Sartori Dias dos Reis¹

Cátedra Sérgio Henrique Ferreira do
Instituto de Estudos Avançados da USP
de Ribeirão Preto

Ribeirão Preto, São Paulo, Brasil

pedro.sartori@usp.br

Rafael Naime Ruggiero²

Departamento de Psicobiologia da
Universidade Federal de São Paulo
(UNIFESP) e Cátedra Sérgio Henrique
Ferreira do Instituto de Estudos
Avançados da USP de Ribeirão Preto

São Paulo, São Paulo, Brasil

rnuggiero@unifesp.br

Mozart Neves Ramos³

Cátedra Sérgio Henrique Ferreira do
Instituto de Estudos Avançados da USP
de Ribeirão Preto

Ribeirão Preto, São Paulo, Brasil

mozartnramos@gmail.com

Resumo

Este trabalho apresenta a *lcde.toolbox*, uma biblioteca desenvolvida em R voltada à análise de dados educacionais. A ferramenta centraliza e automatiza a rotina de processamento, desde o tratamento dos dados e cálculo da Análise de Componentes Principais até a geração de gráficos, criação de mapas de georreferenciamento escolar e montagem de apresentações em *PowerPoint*, padronizando, modularizando e agilizando todo o fluxo analítico. Disponível em código aberto no *GitHub*, a biblioteca está pronta para uso por pesquisadores interessados em conduzir estudos voltados à análise educacional.

Palavras-chave: Linguagem R; Dados Educacionais; Mapas Georreferenciados; Análise de Componentes Principais; Relatórios Automatizados.

Os autores agradecem ao Santander Universidades, à B3 Social e à Fundação Telefônica pelo apoio financeiro ao trabalho da Cátedra Sérgio Henrique Ferreira do IEA da USP de Ribeirão Preto.

1 Introdução

Considerando que a Cátedra Sérgio Henrique Ferreira tem como foco o aprimoramento da aprendizagem dos estudantes e a redução das desigualdades educacionais nas redes públicas parceiras, em consonância com o novo Fundeb (BRASIL, 2020), uma de nossas frentes de atuação é a análise de dados educacionais, com o objetivo de subsidiar decisões e políticas públicas baseadas em evidências. Nesse contexto, adotamos como principal metodologia a Análise de Componentes Principais (ACP) (Mardia, 1979), aplicada a uma matriz composta por cinco indicadores educacionais. A ACP reduz a dimensionalidade dos dados ao gerar duas componentes que, na maioria dos casos, explicam a maior parte da variância total, facilitando a visualização e a interpretação dos resultados.

Para dar suporte à expansão dos trabalhos da Cátedra e garantir maior organização, padronização e agilidade, desenvolvemos a *lcde.toolbox*, uma biblioteca escrita em R (R Project, 2025), cujo nome faz referência ao Laboratório de Ciência de Dados em Educação da Cátedra Sérgio Henrique Ferreira. A biblioteca concentra as principais ferramentas e rotinas utilizadas nas análises de dados educacionais. Ela inclui funções para cálculo e visualização da ACP, geração de tabelas estilizadas e automatização de relatórios em PowerPoint.

Entre suas funcionalidades, destaca-se também a criação de Mapas de Georreferenciamento Escolar (MGE), que permitem representar espacialmente a localização das escolas da rede em um mapa. Essa visualização facilita a identificação de padrões territoriais e regionais relacionados aos dados educacionais. A ferramenta possibilita o cruzamento de diferentes indicadores, como desempenho e nível socioeconômico, por meio de mapas de pontos e mapas de calor, permitindo análises comparativas entre escolas, regiões e redes.

2 Metodologia

A biblioteca *lcde.toolbox* foi desenvolvida na linguagem de programação R, utilizando a IDE (Ambiente de Desenvolvimento Integrado) RStudio. A estrutura do projeto segue o sistema de orientação a objetos S3, escolhido por sua simplicidade e ampla adoção, tanto nos pacotes base do R, como *base* e *stats*, quanto em diversos pacotes disponíveis no CRAN (*Comprehensive R Archive Network*), conforme destacado por Wickham (2015).

O projeto foi organizado em treze classes S3, cada uma responsável por encapsular as propriedades e funcionalidades de diferentes contextos de uso. Essa abordagem foi adotada para garantir a modularidade, a reutilização de código e a clareza na organização da biblioteca. Por exemplo, a classe *pca* concentra os métodos relacionados aos cálculos da Análise de Componentes Principais (ACP), enquanto a classe *pcaviz* é dedicada à visualização dos resultados da ACP, utilizando objetos gerados pela classe *pca*. A Tabela 1 lista as classes e suas respectivas funções:

Tabela 1: **Classes S3 e suas respectivas funções**

Classe	Função principal
colors	Gera paletas de cores para uso em gráficos e mapas
geogg	Cria mapas georreferenciados estáticos com <i>ggplot2</i> ; adiciona camadas de pontos, mapa de calor, contornos, rótulos, legendas e títulos
geoleaf	Cria mapas interativos com <i>leaflet</i> ; adiciona marcadores, mapas de calor, contornos, legendas e <i>pop-ups</i>

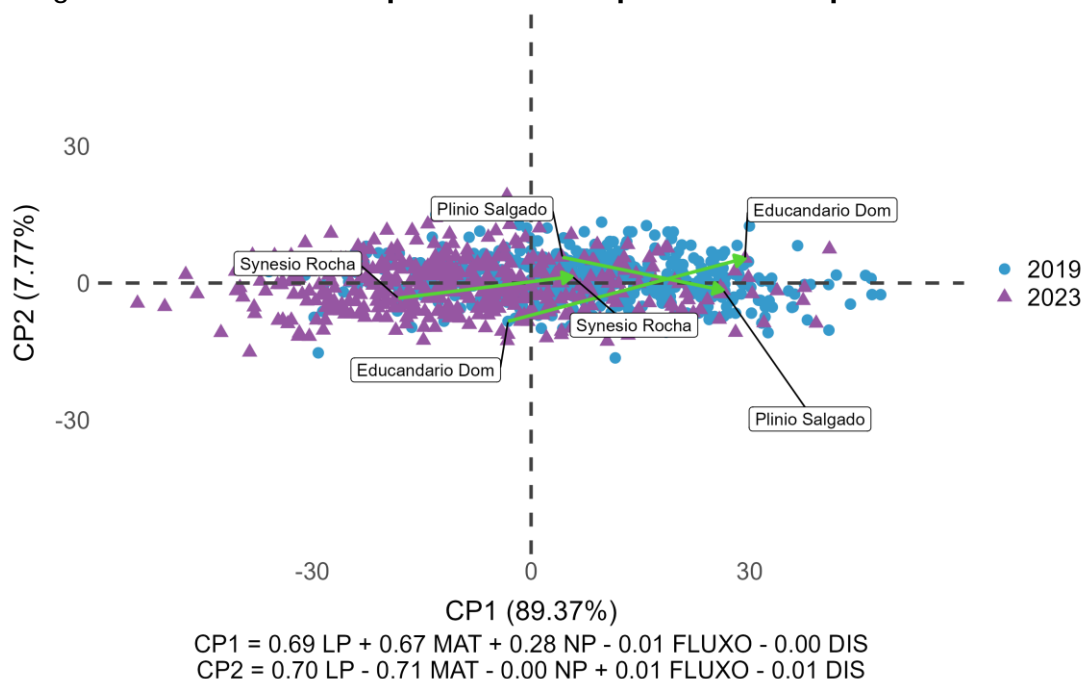
georef	Constrói objeto geoespacial a partir de <i>GeoJSON</i> , de objeto <i>Simple Features</i> (PEBESMA, 2025) ou de vetores de latitude/longitude para a utilização nos mapas georreferenciados; gera camada <i>raster</i> a partir de valores pontuais para a utilização em mapas de calor
ggviz	Converte gráfico <i>ggplot</i> em objeto <i>ggviz</i> para posterior customização; gera gráfico radar (diagrama em formato de teia)
utils	Classe para funções utilitárias, como por exemplo: Abrevia nomes longos de escolas; classifica valores de proficiência em quatro níveis (D–A); associa cada nível a uma cor para a utilização em visualizações; Calcula medidas a partir de matrizes ou <i>Data Frames</i> , magnitude (norma Euclidiana) e magnitude relativa (Utilizadas nos gráficos de radar)
pca	Aplica a análise de componentes principais; Identifica as maiores variações, categoriza escores e filtra resultados
pcaviz	A partir de objeto <i>pca</i> , gera gráfico de dispersão entre dois componentes, gráfico de variações, gráfico de variância explicada e gráfico de cargas; adiciona título e escala de cor
ppt	Automatiza criação de apresentação <i>PowerPoint</i> via <i>officer</i> : carrega template, cria e navega em slides, insere texto, gráficos e tabelas, reordena e salva arquivo
pptpos	Define posições em grade para elementos de slide (linha, coluna, largura, altura e margens); oferece posições pré-prontas
table	Constrói tabela dinâmica estilizada a partir de <i>Data Frame</i> ; exibe variações de PCA, sumário de modelo linear; permite adicionar cabeçalhos, ajustar tamanho e formatação
type	Verifica tipo de objeto (<i>character</i> , <i>numeric</i> , <i>sf</i> , <i>ggplot</i> , etc.) e interrompe execução com erro se o tipo não corresponder ao esperado
vizsize	Cria objeto que define parâmetros de largura, altura, espessura de linha e tamanho de texto para gráficos, garantindo consistência visual

3 Resultados

A Figura 1 apresenta o gráfico de dispersão resultante da ACP aplicada à matriz de dados educacionais da rede municipal de São Paulo, referente aos anos iniciais do ensino fundamental (Reis et al., 2024). O eixo X representa a primeira componente principal (CP1), que explica 89,37% da variância total dos dados, enquanto o eixo Y corresponde à segunda componente (CP2), responsável por 7,77% da variância, que capta as variações residuais não explicadas pela CP1. Pontos azuis indicam a posição das escolas em 2019, e pontos roxos, as mesmas escolas em 2023. Setas verdes destacam aquelas com maior variação ao longo da CP1, componente fortemente influenciada pelos percentuais de aprendizado em língua portuguesa, matemática e pela nota padronizada do Saeb. Abaixo do gráfico, são exibidas as equações das componentes principais, que indicam o peso de cada variável original na composição das componentes. O gráfico foi construído utilizando a classe *pca*, que recebe a matriz de dados e retorna um objeto contendo todas as propriedades da ACP. Ele é passado para a classe *pcaviz*, que utiliza os dados do objeto para a construção do gráfico.

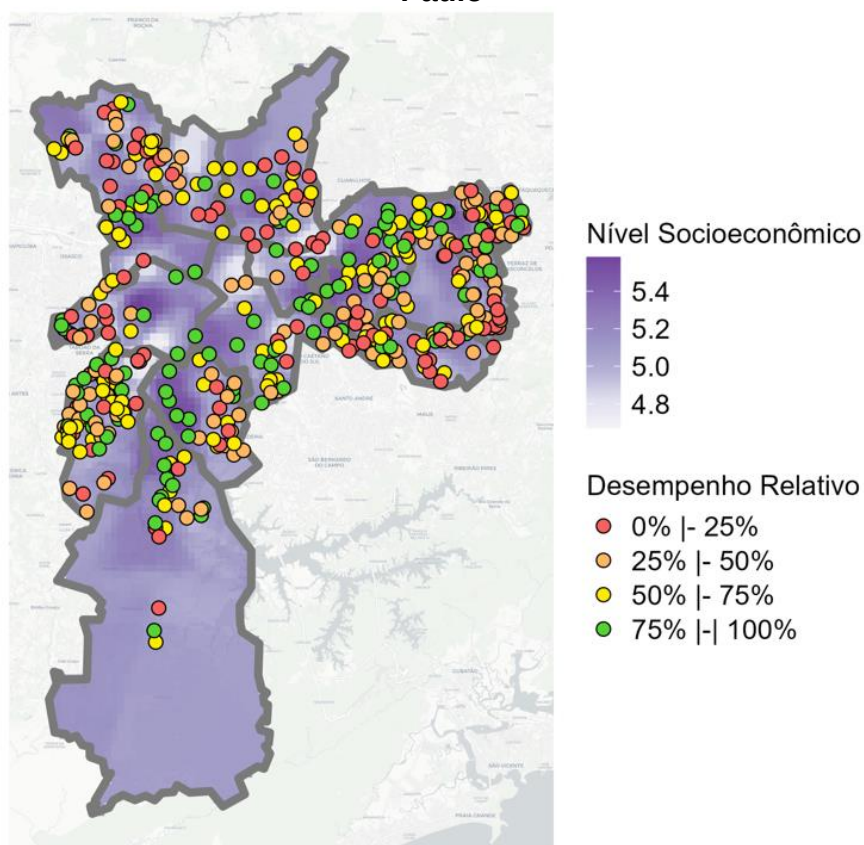
A Figura 2 mostra o Mapa de Georreferenciamento Escolar das escolas municipais dos anos iniciais de São Paulo em 2023, com os contornos das Diretorias Regionais de Educação (DRE). Os pontos vermelhos identificam as escolas com desempenho abaixo do primeiro quartil da primeira componente principal; pontos verdes, aquelas acima do terceiro quartil. A camada roxa ilustra o nível socioeconômico: quanto mais escura a tonalidade, maior o nível. Para gerar esse mapa, foram usados os cinco indicadores da ACP, os dados de nível socioeconômico, as coordenadas das escolas e um arquivo *GeoJSON* com as malhas das DRE's: primeiro, a classe *pca* produziu o objeto com os resultados da ACP, depois, a classe *georef* converteu as malhas em um objeto

Figura 1: Análise de Componentes Principais do Município de São Paulo



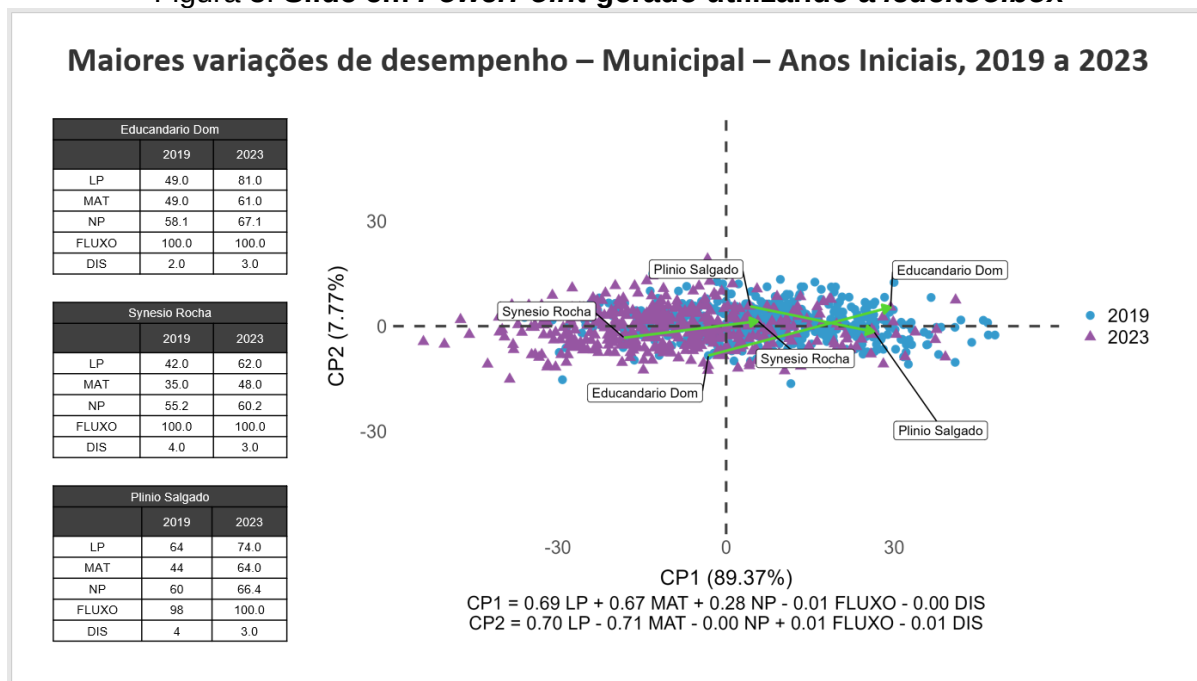
georreferenciado e, por fim, a classe *geogg* combinou esses objetos e os dados socioeconômicos para construir o mapa. Também é possível criar um mapa interativo usando a classe *geoleaf* em vez de *geogg*, permitindo ao usuário navegar, aplicar zoom e consultar informações por meio de *pop-ups*.

Figura 2: Mapa de Georreferenciamento Escolar das escolas municipais de São Paulo



A Figura 3 apresenta um slide gerado com a biblioteca. Além do gráfico da Figura 1, o *layout* exibe, à esquerda, três tabelas com os dados das escolas que registraram as maiores variações na primeira componente principal entre 2019 e 2023. Para criar esse slide, a classe *ppt* foi utilizada para carregar um *template PowerPoint*, abrir um novo *slide*, inserir o gráfico e o título, e posicionar as tabelas utilizando posições predefinidas da classe *pptpos*. As tabelas foram geradas pela classe *table*, que recebe o objeto *pca* e formata automaticamente os registros das três escolas de maior variação.

Figura 3: Slide em *PowerPoint* gerado utilizando a *lcde.toolbox*



4 Conclusões e Considerações Finais

A biblioteca *lcde.toolbox* demonstrou ser uma ferramenta eficaz para apoiar as análises educacionais, promovendo agilidade na geração de resultados e garantindo a escalabilidade dos projetos desenvolvidos. A automação proporcionada pela biblioteca simplificou a criação de relatórios de Análise de Componentes Principais e Mapas de Georreferenciamento Escolar, que podem ser gerados com apenas alguns cliques, mesmo em redes de grande porte, como a municipal de São Paulo apresentada neste artigo. Isso permitiu redirecionar o tempo antes dedicado a tarefas repetitivas, como a produção de gráficos e a montagem de apresentações, para a análise dos resultados. Como a biblioteca está disponível publicamente no *GitHub*, ela também pode ser utilizada por pesquisadores externos interessados em realizar análises educacionais semelhantes de maneira independente.

5 Referências

BRASIL. Ministério da Educação. Portaria n. 931, de 21 de março de 2005. Institui o Sistema de Avaliação da Educação Básica – Saeb. Diário Oficial da União: seção 1, Brasília, DF, n. 55, p. 16-17, 22 mar. 2005.

BRASIL. Ministério da Educação. Decreto n. 6.094, de 24 de abril de 2007. Diário Oficial da União: seção 1, Brasília, DF, n. 79, p. 5, 24 abr. 2007. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/decreto/d6094.htm.

BRASIL, Ministério da Educação, Lei nº 14.113, de 25 de dezembro de 2020, que trata da regulamentação do Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (Fundeb), art. 212-A da Constituição Federal, (2020).

BRASIL, Ministério da Educação, Portaria nº 975, de 13 de dezembro de 2022, que aprova a metodologia de cálculo dos indicadores de atendimento e de melhoria de aprendizagem, com redução de desigualdades, consoante o art. 5º, inciso III, o art. 14, caput e §§ 2º e 3º, e o art. 15, inciso III, da Lei nº 14.113, de 25 de dezembro de 2020, que regulamenta o Fundeb, para o exercício financeiro de 2023.

<https://www.gov.br/fnde/pt-br/aceso-a-informacao/acoes-e-programas/financiamento/fundeb/legislacao/2022/portaria-no-975-de-13-de-dezembro-de-2022/view>

BRASIL, Ministério da Educação, Redes de ensino inabilitadas à complementação VAAR 2024 pelo não cumprimento de condicionalidades de melhoria de gestão previstas no art. 14, § 1º, incisos I a V, da Lei nº 14.113/2020, ou por não terem alcançado evolução em indicadores de atendimento e de melhoria da aprendizagem com redução das desigualdades previstos no art. 14, § 2º da Lei nº 14.113/2020. Disponível em:

<https://www.gov.br/fnde/pt-br/aceso-a-informacao/acoes-e-programas/financiamento/fundeb/2024/Redesinabilitadaspormotivo.pdf>

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M.; Multivariate Analysis, Academic Press, London, 1979. Disponível em:

<https://statisticalsupportandresearch.files.wordpress.com/2017/06/k-v-mardia-j-t-kent-j-m-bibby-multivariate-analysis-probability-and-mathematical-statistics-academic-press-inc-1979.pdf>.

PEBESMA, Edzer. Simple Features for R. Disponível em: <https://r-spatial.github.io/sf/articles/sf1.html>. Acesso em: 06 de maio de 2025.

RAMOS, M.N.; SILVA, L.; SILVA, J.B.P.; FILHO, A.J.C.; “Uma análise longitudinal do Ideb em municípios de médio porte do estado de São Paulo”, Ensaio: aval. pol. públ. Educ., Rio de Janeiro, v.31, n.118, p. 1-22, jan./mar. 2023. Disponível em: <https://doi.org/10.1590/S0104-40362023003103818>

RAMOS, M.N.; FILHO, A.J.C.; SILVA, J.B.P.; NASCIMENTO, E.F.V.B.C. 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. Disponível em: <https://doi.org/10.1590/S0104-40362021002903286>

REIS, P.S.D.; JÚNIOR, J.H.R.; RUGGIERO, R.N.; RAMOS, M.N.; “Uso de modelagens estatísticas e de ciências dos dados para reduzir as desigualdades escolares da rede pública municipal do Município de São Paulo”, IX Congresso Brasileiro de Teoria de Resposta ao Item (TRI) e Métodos Quantitativos em Avaliação – IX CONBRATRI, Rio de Janeiro (RJ), 2024.

R FOUNDATION. The R Project for Statistical Computing. Disponível em: <https://www.r-project.org/>. Acesso em: 06 de maio de 2025.

WICKHAM, Hadley. Advanced R. Boca Raton: CRC Press, 2015. eBook. ISBN 978-1-4665-8697-0.