"Planeta Água: a cultura oceânica para enfrentar as mudanças climáticas no meu território"
SECT - ICE
XII Semana de Ciência e Tecnologia SECT ICE
20 a 23 de Outubro de 2025
Realização:

AmazonasDataHub: Uma Biblioteca de Dados em R

[1]Prof. Dr. Leonardo Brandão Freitas do Nascimento (Orientador), [2]Nelson Geraldo Aquino de Carvalho (Pesquisador).

Universidade Federal do Amazonas , Departamento de Estatística. Av. Rodrigo Otávio Jordão Ramos, 6200, Coroado I, 69080-900, Manaus AM, Brasil.

[1]nascimento@ufam.edu.br; [2]nelson.carvalhoac@gmail.com

Palavras-Chave: Amazonas; Biblioteca; Dados; Pacote; R.

Introdução

Com a alta demanda por profissionais em estatística e ciência de dados, a criação de materiais de apoio, que favoreçam um processo de aprendizado rápido e ágil, torna-se necessária. Nesse cenário, a disponibilização de bases de dados tratadas tem se mostrado fundamental, pois possibilita que estudantes e pesquisadores tenham acesso a dados reais e organizados, com os quais podem reproduzir estudos, resultados e diferentes tipos de análises. Essa viabilização contribui para fins didáticos, tanto de ensino quanto em conhecimento.

Os pacotes e bibliotecas em R desempenham um papel importante na disponibilização de conjuntos de dados. Nesse sentido, o presente trabalho propõe a criação de um data hub — uma plataforma voltada à centralização de informações estruturadas. Esse recurso, de forma digital e acessível, permitirá ampliar o acesso a bases de dados sobre o Estado do Amazonas. Através da linguagem de programação R (R Core Team, 2024), o projeto Amazonas DataHub, visa a criação do pacote "amazonasdatahub". Tal pacote permitirá o armazenamento, organização, consolidação e centralização das bases de dados utilizadas em pesquisas científicas e oriundas de informações registradas do Estado, provenientes de sites e relatórios. O desenvolvimento do projeto seguiu padronizações para manipulação de bases de dados propostas por Wickham, Rundel e Grolemund (2023) e para a criação e disponibilização do pacote propostas por Wickham e Bryan (2023). Portanto, esse pacote permitirá o acesso à informação e a conjuntos de dados bem estruturados, viabilizando práticas de estudos, ensino e aplicações de métodos estatísticos, como análise exploratória de dados, mineração de dados, análise de regressão, entre outros, além de propiciar a divulgação de resultados científicos e a utilização de exemplos contextualizados.

Material e Métodos

Para o desenvolvimento do pacote na linguagem de programação R, foram utilizadas as padronizações propostas por Wickham e Bryan (2023), que utilizam os as bibliotecas *devtools* e *usethis*. Além disso, foram utilizadas ferramentas que permitem aplicar a ciência de dados de forma eficiente e reprodutível. Tais padronizações são detalhadas no livro de Wickham, Rundel e Grolemund (2023) e Irizarry (2019).

O pacote *devtools*, que é um conjunto de pacotes que auxiliam em diversos aspectos do desenvolvimento precisa ser instalado no R com *install.packages("devtools")* e carregado com *library(devtools)*. Para inicializar um novo pacote, utiliza-se o comando *usethis::create_package("amazonasdatahub")*.

Após importar os dados, a próxima etapa é a realização da limpeza e manipulação. Essa etapa é importante porque a estrutura consistente das bases de dados permite uma melhor análise das mesmas. Funções do *dplyr* foram usadas para ajustar e padronizar os nomes das colunas para o padrão *camel-case*.

Usamos a função *pivot_longer*, do pacote *tidyr* para aplicar a transformação para o formato tidy, onde cada coluna é uma variável e cada linha é uma observação (WICKHAM; VAUGHAN; GIRLICH, 2024). Esse processo é realizado, primeiramente, pela importação das bases no ambiente de desenvolvimento, e em seguida, é utilizado o método *usethis::use_data*. No diretório "/R" do projeto, foram criados os códigos com modelos roxygen2, nos quais as documentações das funções foram escritas. Em seguida, utilizou-se a função *devtools::document()* para gerar e salvar cada documentação. Para a documentação geral do pacote, empregou-se a função *usethis::use_readme_rmd()*, que cria um arquivo no formato RMarkdown, é possível realizar a instalação do pacote diretamente do GitHub através do comando

devtools::install_github("onelsoncarvalho/amazonasdatahub)

ou

remotes::install_github("onelsoncarvalho/amazonasdatahub").

Resultados e Discussão

Por meio das aplicações das metodologias discutidas na seção de Material e Métodos, foi possível desenvolver um pacote estável e escalável, permitindo tanto a reprodutibilidade quanto a inserção de novos conjuntos de dados. A Tabela 1 apresenta as bases de dados reunidas e documentadas até o momento.

Tabela 1 - Bases de Dados disponíveis atualmente no amazonasdatahub

Base de Dados	Descrição
agriculture_amazonas	Plant production data from the Institute of Sustainable Agricultural and Forestry Development of Amazonas - IDAM (2023)
aids_amazonas	A dataset of AIDS occurrences from 2011 to 2023 in Amazonas
humidity_manaus	Time series with Extreme Values of Limited Range

malaria_amazonas	An integrated dataset of malaria notifications in the state of Amazonas
gdp_amazonas	Gross Domestic Product - GDP of Amazonas
rionegro_amazonas	A dataset of the Rio Negro River (Amazonas) level from 2003 to 2023
srl_muni	Physical literacy and reading performance of amazonian schoolchildren: an association study

A seguir, temos algumas aplicações gráficas dos dados providos. Na Figura 1, temos a aplicação de um diagrama de dispersão indicando uma correlação linear positiva dos dados de área plantada e área colhida de mandioca da base de dados agriculture_amazonas. A série temporal da Figura 2 é uma das diversas aplicações possíveis para o conjunto aids_amazonas, e mostra a incidência de casos de AIDS agrupadas por gênero, de 2011 a 2023 filtrados para Manaus, mostrando os aumentos e reduções de casos registrados. Já a Figura 3 é uma matriz de correlação feita utilizando o método de Spearman aplicada na base de dados srl_muni, mostrando a correlação das variáveis, como a de Competência Motora (mc) e a pontuação de desempenho em leitura (tde), que apresentam uma correlação significante.

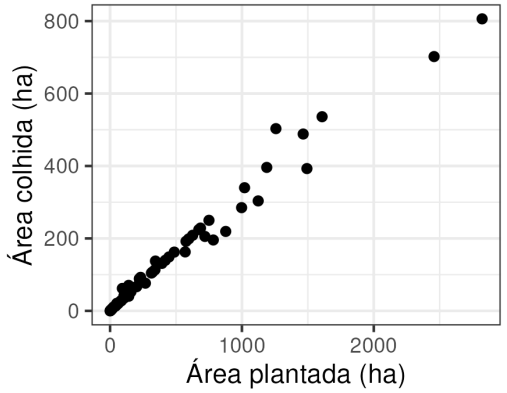

Figura 1. Diagrama de dispersão: área plantada e área colhida da produção de mandioca nas unidades locais assistidas pelo IDAM

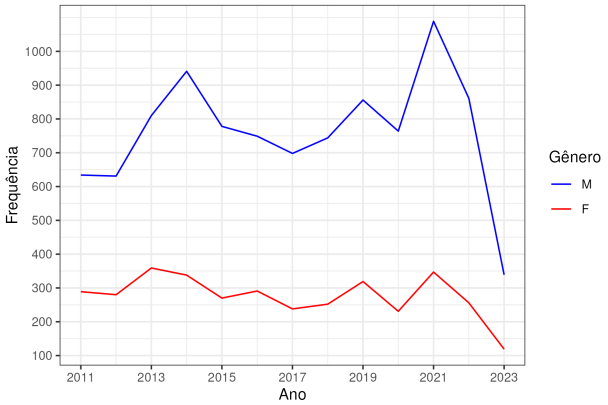

Figura 2. Contagem de casos de AIDS registrados em Manaus entre 2011 a 2023

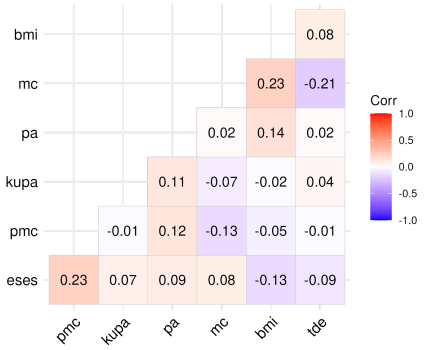

Figura 3. Matriz de correlação das variáveis do Physical Literacy e Desempenho em Leitura

Como parte do projeto, o website desenvolvido exibe mais informações sobre o Amazonas DataHub e das bases de dados (documentação, fontes e exemplos) e pode ser acessado por: <https://onelsoncarvalho.github.io/amazonasdatahubsite/> (AMAZONAS DATAHUB, 2025).

Conclusões

Na versão atual, a biblioteca/pacote amazonasdatahub disponibiliza sete bases de dados totalmente documentadas, classificadas como climáticas, de saúde, econômicas, agrícolas e de ciências sociais. O projeto foi desenvolvido de forma escalável, permitindo a inserção de novos conjuntos de dados. Isso é possível devido a utilização das metodologias propostas por Wickham e Bryan (2023). Tais propostas possibilitam que o data hub desenvolvido ofereça dados estruturados, aplicáveis em análises estatísticas diversas. O formato estruturado das bases de dados oferecidas propicia a criação de modelos estatísticos, modelos de aprendizado de máquina, bem como a reprodutibilidade, permitindo que estudos e análises replicados, comparados e validados por diferentes usuários. Por fim, o Amazonas DataHub contribui para o letramento estatístico ao disponibilizar dados científicos de forma acessível, possibilitando o uso de dados regionais no ensino, e assim, contribuindo para o aprendizado dos conceitos estudados e colaborando para a formação de profissionais em um cenário de alta demanda de estatísticos e cientistas de dados.

Referências

[1] AMAZONAS DATAHUB. *Amazonas DataHub: dados do Amazonas para aplicação de métodos estatísticos*. Disponível em: <https://onelsoncarvalho.github.io/amazonasdatahubsite/>.

[2] IZARRY, R. A. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*. Disponível em: <https://doi.org/10.1201/9780429341830>.

[3] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2024. Disponível em: <https://www.R-project.org/>.

[4] WICKHAM, Hadley; RUNDEL, Mine Çetinkaya; GROLEMUND, Garrett. *R for Data Science (2e)*. 2. ed. O'Reilly Media, 2023. Disponível em: <https://r4ds.hadley.nz>. Acesso em: 12 jan. 2025.

[5] WICKHAM, Hadley; BRYAN, Jennifer. *R Packages. 2. ed.* O'Reilly Media, 2023. Disponível em: <https://r-pkgs.org>. Acesso em: 12 jan. 2025.

[6] WICKHAM, Hadley.; VAUGHAN, Davis; GIRLICH, Maximilian. *tidyr: Tidy Messy Data*. Disponível em: <https://cran.r-project.org/package=tidyr>. Acesso em: 10 jan. 2025.

[7] WICKHAM, Hadley, FRANÇOIS ROMAIN, HENRY, Lionel, MÜLLER, Kirill, VAUGHAN, Davis. (2025). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4,. Disponível em: < https://dplyr.tidyverse.org>. Acesso em 10 jan. 2025.