



Reprodutibilidade e compartilhamento de resultados na pesquisa científica

MILZ, Beatriz¹

Resumo: Este texto explora a importância da reprodutibilidade na pesquisa científica, apresentando o fluxo de trabalho convencionalmente utilizado em pesquisas na Ciência Ambiental, e uma alternativa proporcionada pelo uso de ferramentas baseadas no paradigma de programação letrada, como o Quarto e o RMarkdown.

Palavras-chave: pesquisa reprodutível, Quarto, análise de dados, programação letrada.

1 Introdução

Ao realizar uma pesquisa, uma importante etapa é o compartilhamento dos resultados, seja através de documentos para comunicação com pares (como artigos científicos, trabalhos de conclusão de curso, dissertações, teses, apresentações em congressos, entre outros) ou para comunicação com a sociedade (como relatórios, apresentações, textos, vídeos, entre outros). Para tanto, é importante que a escolha do formato de comunicação esteja alinhada ao público-alvo, assim como a possibilidade de gerar diferentes tipos de documentos para comunicação científica.

O objetivo deste texto é discutir sobre o fluxo de trabalho mais comum utilizado atualmente, e apresentar a possibilidade de um

¹ Doutora em Ciência Ambiental pelo PROCAM, do Instituto de Energia e Ambiente da Universidade de São Paulo (IEE/USP). milz.bea@gmail.com



fluxo de trabalho que facilita o compartilhamento de resultados de pesquisa posteriormente, em diferentes tipos de documentos.

Esse texto está fundamentado nas experiências acadêmicas da autora nos últimos cinco anos, em sua atuação na área da Ciência Ambiental.

2 Desenvolvimento

2.1 Fluxo de trabalho convencional

Em pesquisas na área ambiental, existem muitos elementos criados a partir de análise de dados, como: tabelas, gráficos, mapas, resultados de modelos, entre outros. Esses elementos podem ser elaborados utilizando diferentes softwares, dependendo do resultado a ser criado e por escolhas da equipe de pesquisa.

Os softwares mais comuns são de planilhas eletrônicas (como Microsoft Excel, Apache OpenOffice Calc, Google Sheets), criação de mapas (como o QGis e o ArcGis), e softwares estatísticos (SPSS, JASP, Jamovi, entre outros). Um fluxo de trabalho comum é elaborar a análise em algum destes softwares, e posteriormente copiar o resultado e colar em um software de edição de texto (como por exemplo, o Microsoft Word, Apache OpenOffice Writer, Google Docs, entre outros).

Existem algumas desvantagens com esse fluxo de trabalho. Por ser manual, qualquer alteração ou atualização nos dados de pesquisa implica em realizar novamente as análises, despendendo um longo tempo para a realização da tarefa. Outra



desvantagem é a falta de reprodutibilidade, que será discutida em seguida.

2.2 Qual é a importância da reprodutibilidade?

A reprodutibilidade é um pilar fundamental da pesquisa científica, permitindo que experimentos, análises e conclusões sejam verificados e validados por pares, ou seja, outras pessoas que fazem pesquisa na mesma área. Além de promover a confiança na ciência, a reprodutibilidade facilita a colaboração entre pesquisadores, permitindo que análises sejam revisadas, aprimoradas e adaptadas para outros contextos. A reprodutibilidade não se limita apenas à análise de dados; ela também envolve a documentação adequada de métodos, disponibilização da base de dados original, entre outros.

Baseado em uma pesquisa com mais de 1.500 cientistas realizada pela Revista Nature, Baker (2016) apresenta importantes resultados sobre a reprodutibilidade, como "mais de 70% dos pesquisadores tentaram e falharam em reproduzir os experimentos de outro cientista", e "mais da metade não conseguiu reproduzir seus próprios experimentos" (Baker, 2016, p. 452).

Os principais fatores apontados pelos cientistas respondentes, para impossibilitar a reprodutibilidade de pesquisas, são: compartilhamento seletivo de resultados (*selective reporting*) pressão para publicação de artigos, baixo poder estatístico, análises de baixa qualidade, orientação insuficiente, replicação insuficiente no laboratório original,



métodos não disponíveis, código utilizado na análise não disponível, dados brutos não disponíveis, entre outros (Baker, 2016).

2.3 Uma alternativa para as ferramentas utilizadas

Uma alternativa ao fluxo convencional de trabalho apresentado passa pela discussão das ferramentas e softwares utilizados.

Os softwares para análise de dados apresentados anteriormente apresentam limitações em termos de reprodutibilidade. Uma alternativa é a realização das análises utilizando linguagens de programação (por exemplo R ou Python). Entretanto, essa forma ainda não é amplamente utilizada nos cursos, laboratórios e centros de pesquisa na área ambiental no Brasil.

A utilização de linguagens de programação para analisar dados possibilita o aumento da reprodutibilidade. Nessa perspectiva, cada etapa de análise é documentada em forma de código, que pode ser compartilhada com pares para que seja consultada, avaliada, e aprimorada em pesquisas futuras.

Realizar as análises de dados utilizando linguagens de programação é um avanço em termos de reprodutibilidade. No entanto, é possível ir além da simples programação, através da aplicação do conceito de programação letrada (*Literate programming*) (KNUTH, 1992).

A programação letrada é um conceito que surgiu com Donald Knuth na década de 80. Esse paradigma de programação consiste em colocar o código fonte de um programa e texto de



sua documentação em um único arquivo, de forma que pessoas possam ler e entender o conteúdo. Donald Knuth também desenvolveu o TeX: um sistema tipográfico que permite produzir documentos com alta qualidade.

2.4 Ferramentas baseadas em programação letrada

O paradigma da programação letrada é amplamente utilizado na ciência de dados, possibilitando a criação de documentos reprodutíveis.

O RMarkdown (XIE, ALLAIRE & GROLEMUND, 2018; ALLAIRE *et al*, 2024) é um pacote que permite a criação de relatórios reprodutíveis utilizando a linguagem R (R CORE TEAM, 2022), lançado em 2014 e mantido pela equipe da empresa Posit² (anteriormente chamada de RStudio). Com os avanços e aprendizados obtidos, a equipe da Posit desenvolveu e lançou a nova geração do RMarkdown em 2022, chamada: Quarto.

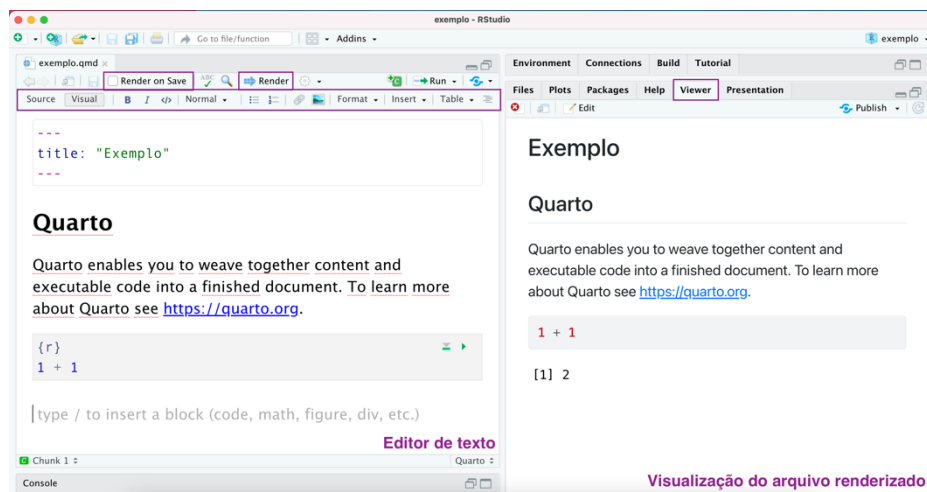
O Quarto (ALLAIRE, 2023) é um software de código aberto para escrita técnica e científica. Um importante avanço do Quarto, se comparado ao RMarkdown, é não ser dependente do R, podendo ser utilizado com outras linguagens, principalmente Python, Julia e JavaScript.

A Figura 1 apresenta uma captura de tela da Interface de Desenvolvimento Integrada (IDE) RStudio, com um arquivo simples feito com Quarto à esquerda (chamado de *exemplo.qmd*), e a visualização renderizada a partir deste arquivo à direita.

² Posit - <https://posit.co/>



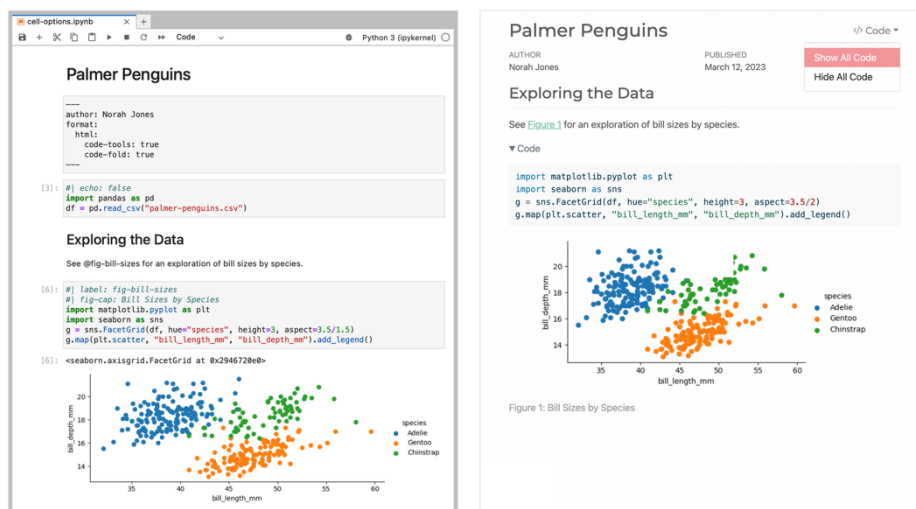
Figura 1. Captura de tela: Exemplo de um arquivo Quarto (.qmd),



Fonte: Elaborado pela autora.

É possível utilizar o Quarto combinado ao Jupyter Notebook, uma ferramenta muito utilizada em ciência de dados, principalmente pela comunidade de pessoas usuárias de Python. Na Figura 2, é possível ver um exemplo dessa combinação do uso do Quarto com o Jupyter Notebook, onde ao lado esquerdo é possível ver o código, e ao lado direito é possível ver a versão renderizada a partir deste código.

Figura 2. Exemplo de uso do Quarto com um Jupyter Notebook



Fonte: Documentação do Quarto (<https://quarto.org/>).

A adoção do software Quarto permite não apenas melhorar a reprodutibilidade nas pesquisas, mas também possibilita a flexibilidade de gerar diferentes tipos de arquivos, como apresentações, livros, artigos, relatórios em PDF, word, e HTML, entre outros.

Essas ferramentas representam uma evolução significativa em relação aos fluxos de trabalho convencionais, integrando análise de dados e criação de documentos em um único processo. Com elas, é possível gerar documentos que apresentam os textos, códigos, resultados de análises, gráficos, tabelas, mapas, facilitando a atualização de análises e a geração de múltiplos formatos de saída a partir de um único arquivo.



3 Desafios

Apesar de apresentar benefícios claros, a adoção dessas ferramentas e fluxos de trabalho baseados em programação letrada apresentam desafios. Estes desafios podem ser tanto técnicos quanto culturais, refletindo a complexidade de mudar paradigmas estabelecidos na pesquisa científica.

Dentre eles, destaca-se a curva de aprendizagem. O processo de aprendizado pode ser longo, especialmente para pessoas pesquisadoras que não têm experiência prévia com programação.

Algumas universidades já oferecem disciplinas sobre o tema, porém essa adesão ainda não é ampla, principalmente se tratando em cursos voltados para a Ciência Ambiental.

Além disso, a colaboração com pesquisadoras e pesquisadores que utilizam o fluxo convencional de trabalho, pode ser dificultosa e gerar retrabalho para a pessoa responsável sozinha em estruturar o documento final.

4 Conclusão

Para superar esses desafios, é crucial promover uma cultura de valorização da pesquisa reprodutível dentro da comunidade científica.

Isso inclui o desenvolvimento de políticas institucionais que incentivem práticas de pesquisa reprodutível e o fornecimento de disciplinas, recursos e treinamentos adequados sobre o tema.



Além disso, é importante destacar os benefícios a longo prazo da reprodutibilidade, não apenas para a integridade da pesquisa, mas também para a carreira das pessoas pesquisadoras. Pesquisas reprodutíveis podem ser validadas e replicadas em outros contextos, ampliando o alcance e a contribuição da pesquisa original.

A adoção de ferramentas e fluxo de trabalho que permita um desenvolvimento de pesquisas mais reprodutíveis possibilitará que a comunidade científica caminhe em direção para ter uma maior transparência e colaboração nas pesquisas.

Referências

ALLAIRE, J. *quarto: R Interface to 'Quarto' Markdown Publishing System*. Versão do pacote R 1.3, 2023. Disponível em: <https://CRAN.R-project.org/package=quarto>. Acesso em: 23 de março de 2024.

ALLAIRE, J.; XIE, Y.; DERVIEUX, C.; MCPHERSON, J.; LURASCHI, J.; USHEY, K.; ATKINS, A.; WICKHAM, H.; CHENG, J.; CHANG, W.; IANNONE, R. *rmarkdown: Dynamic Documents for R*. Versão do pacote R 2.26, 2024. Disponível em: <https://github.com/rstudio/rmarkdown>. Acesso em: 23 de março de 2024.

BAKER, M. 1.500 scientists lift the lid on reproducibility. *Nature*, v. 533, p. 452-454, 2016. Disponível em: <https://doi.org/10.1038/533452a>. Acesso em: 23 de março de 2024.

KNUTH, D. *Literate Programming*. Stanford: Center for the Study of Language and Information, 1992.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Viena, Áustria: R Foundation for Statistical Computing, 2022. Disponível em: <https://www.R-project.org/>. Acesso em: 23 de março de 2024.



XIE, Y.; ALLAIRE, J.; GROLEMUND, G. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman and Hall/CRC, 2018. ISBN 9781138359338. Disponível em: <https://bookdown.org/yihui/rmarkdown>. Acesso em: 23 de março de 2024.