EDITORIAL

Open-Source Statistical Software: Revolutionising Health Research

Masood Ali Shaikh, Syed Muhammad Mubeen

Data are everywhere, and medical and public health research depends on making sense of it. This requires utilisation of advanced statistical and geographical information system (GIS) tools to glean insights from data, guide health and public policies, and promote health sciences to reduce the global burden of morbidity and mortality. Proprietary statistical software programmes like SAS, SPSS, and Stata, along with ArcGIS for mapping and spatial analysis, have dominated the statistical analysis field for decades. The emergence of open-source software for statistical analysis, such as R, Python, and Julia, as well as spatial analysis software like QGIS, is challenging these proprietary behemoths. These open-source software solutions provide cutting-edge analytical capabilities at no financial cost, thereby democratising access to health data analysis and insights for improving the health of populations globally.

The limited availability of health research funding in resource-constrained developing countries like Pakistan, makes the provision of free, easily accessible offline quality data analysis programmes, free from any licensing fees, a fundamental change.

The hallmark of open-source software is its unrestricted access and availability for use, together with publicly accessible source code; enabling users to examine, modify, and disseminate without any limitations. The major open-source statistical software include Python, R, Julia for statistical analysis, and QGIS for spatial analysis and mapping. Each has numerous global online communities that strive to support and answer user's questions and inquiries. Advocating for the appropriate application of statistical tests and software without the burden of exorbitant initial and annual maintenance licensing fees. These open-source software are also free from restrictive usage agreements like the use of the software on the limited number of computers or on multiple computers simultaneously.[1]

R, Python, Julia and QGIS have intuitive and user-friendly syntax complemented by packages, libraries and plugins that supplement the capability of the core application. Python is a leading open-source software for machine learning and data science.[2] R, albeit developed primarily for statistical computing, has an extensive range of libraries for data entry, manipulation, visualisation and modelling.[3] Likewise, Julia, is a high-level programming language designed for statistical analysis.[4] QGIS software, an open-source Geographic Information System, provides extensive spatial analysis and mapping capabilities for geographic data, comparable to those of expensive proprietary GIS software.[5] Python, R, and Julia also have GIS and spatial analysis capabilities.

More specifically, R provides a plethora of packages for data scientists, including data handling (tidyverse), data visualization (ggplot2), interactive plots (plotly), web interactive dashboards (shiny), geographic data visualisation (tmap), creating interactive tables (DT), unified interface for machine learning (caret, mlr3), and deep learning (keras). Similarly, Python is replete with packages for data scientists: for data manipulation and analysis (Pandas), numerical computing (NumPy), data visualization (Matplotlib), statistical modelling (Statmodels), machine learning (scikit-learn), and deep learning (TensorFlow & PyTorch). Likewise, Julia has packages for managing and manipulating data frames (DataFrames.jl), as well as for defining, fitting, and evaluating various statistical models (StatsModels.jl), machine learning (SciML.jl), and deep learning (Flux.jl). These packages/libraries represent a few among the thousands available for each of these three software applications. QGIS extends these capabilities to spatial data by providing plugins for creating interactive plots (DataPlotly), geocoding and data conversion tasks (MMQGIS), exporting QGIS projects to web maps (GIS2Web), generating 3D models of urban areas (3D City Builder), and creating 3D visualisations (QGIS2threejs), among others. The wealth of available statistical analysis options facilitates a common analysis language and platforms across disciplines and is instrumental for promoting collaborative interdisciplinary strategies to address health morbidity and mortality burden.[6]

These open-source software applications have greatly improved the fields of statistical and spatial analysis for researchers across several scientific disciplines in low and

[1]Department of Medicine, College of Medicine, Korea University, South Korea, [2]Hamdard College of Medicine & Dentistry,Hamdard University, Karachi, Pakistan

Correspondence: Syed Muhammad Mubeen.

Email: dr_mubeen@hotmail.com

middle-income countries. Scientific integrity depends on the reproducibility of results and the dissemination of data. Open-source statistical and spatial analysis software enhance transparency and reproducibility of reported results by facilitating the sharing of analysis scripts with data. Therefore, promoting better peer review, replication, and verification of reported findings.[7] For more effective peer review and transparency in scientific research, Python's Jupyter Notebooks offer the capability to integrate data, and analysis scripts with comments/explanations within a single interactive and shareable document.[8] Online systems such as GitHub (https://github.com/) allow complimentary hosting of data, and analysis scripts utilized in published articles based on R, Python, Julia and QGIS. While the capacity to handle large spatial data makes QGIS particularly useful for studying health geography and spatial epidemiology.[9]

Open-source software in conjunction with platforms like GitHub promotes openness and reproducibility of results in addition to serving as an excellent resource for education and learning, for new and established health researchers, especially in developing countries. In Pakistan, proprietary statistical packages dominate, as the majority of published research articles in internationally, regionally, and nationally recognized indexed medical journals primarily use them for the statistical analysis. For some, or many, there is no choice but to use the pirated/cracked i.e. unauthorized copies of software that are sold at the fraction of the cost, without permission from the copyright holders. These pirated versions may also carry the risk of malware or viruses that can damage computers or, even worse, compromise data integrity.

There is a cost – albeit non-financial – associated with the use of open-source programmes. Transitioning from proprietary software to open-source alternatives may involve a steep learning curve. The absence of prompt technical support associated with licensed proprietary software is another issue, which is obviously not available to the users of pirated software. But there are websites like 'Stack Overflow' (https://stackoverflow.com/) where users can freely ask, and often get fairly quick answers, to technical questions about the use and interpretation of outputs from these open-source statistical and GIS programmes. Switching to open-source software for statistical analysis addresses issues of legality, ethics, cost, and security, while promoting scientific reproducibility and ensuring professional integrity in the conduct of health research.

https://doi.org/10.47391/JPMA.25-101

Disclaimer: None.

Conflict of Interest: None.

Source of Funding: None.

References

1. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015 26;348:1422-5.
2. McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 2010, 51-56.
3. Wickham H, Grolemund G. R for Data Science. 2016, O'Reilly Media.
4. Bezanson J, Edelman A, Karpinski S, Shah VB. Julia: A fresh approach to numerical computing. SIAM review. 2017;59:65-98.
5. Bivand RS, Pebesma EJ, Gómez-Rubio V, Pebesma EJ. Applied spatial data analysis with R. New York: Springer; 2008.
6. Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nature human behaviour. 2017 10;1:1-9.
7. Peng RD. Reproducible research in computational science.Science.2011;334:1226-7.
8. Van Rossum G, Drake FL. Python 3 Reference Manual. 2009 CreateSpace Independent Publishing Platform.
9. Dunnington D, Lovelace R, Sullivan C. Geocomputation with R. 2019. RC Press.