REGION

The Journal of ERSA
Powered by WU

Volume 7, Number 3, 2020, E1–E5
DOI: 10.18335/region.v7i3.357

journal homepage: region.ersa.org

The Potential of Notebooks for Scientific Publication: Reproducibility, and Dissemination

Francisco Rowe[1], Gunther Maier[2], Daniel Arribas-Bel[1] and Sergio J. Rey[3]

[1] University of Liverpool, Liverpool, UK
[2] Modul University, Vienna, Austria
[3] University of California, Riverside CA, USA

Received: 25 December 2020/Accepted: 25 December 2020

1 Background

Recent developments and discussions concerning the SARS-Cov-2 virus and the development of a vaccine illustrate once again the necessity to assume "that scientific claims are supported by solid evidence" (Branco et al. 2017). In recent years, however, we see increasing evidence that casts doubts on this assumption (e.g. Fanelli 2009, Ioannidis 2011, Prinz et al. 2011, Begley, Ellis 2012, Fokkens et al. 2013, Open Science Collaboration 2015). "As a result, there is an increasingly urgent call for validation and verification of published research results, both within the academic community and the public at large (e.g. Naik 2011, Zimmer 2012, Begley 2012, Editorial 2013a,b, Branco 2012)" (Branco et al. 2017, p. 1). This is particularly important at a time when the scale and complexity of scientific studies grow, and replicability and reproducibility of scientific research has gained salience (Peng 2011).

Recent developments in software and web browser technology may help in solving this problem. They have enabled exciting and fast-moving developments in many areas of research, among them tools that can help validate and verify research as well as stimulate knowledge transfer among researchers. Computational notebooks, particularly Jupyter Notebooks, represent a major advance for scientific research. A Jupyter Notebook is an open-source web application which enables creating and sharing documents containing live code, equations, visualisations and narrative text (Jupyter Project 2019). A Jupyter notebook comprises a series of 'cells' containing executable code, or markdown, along with the popular HTML markup language for prose descriptions and LaTeX for mathematical equation write up. Jupyter Notebooks have enabled a new type of programming which emphasises a prose-first approach where exposition with human-friendly narrative is threaded with code blocks. The Jupyter Notebook was originally developed by Fernando Perez and Brian Granger in the Python programme language in 2011 and known as IPython Notebooks. In 2013, the technology was expanded to allow for additional programming languages and renamed 'Jupyter'[1].

The interactive and narrative nature of computational notebooks provide unique opportunities for sharing computational results, enabling reproducibility and publishing scientific research. Traditionally, code, data, results, and their exposition are stored in separate files, which is a source of disconnect and easy misalignment. Computational

[1] 'Jupyter' is an acronym for Julia, Python and R, three of the main modern languages for scientific computing.

notebooks allow conducting analyses and integrating code, results and descriptive text into a single 'computational narrative' to be shared, read and executed by others (Pérez, Granger 2015, Kluyver et al. 2016). Attracted by the ability to combine executable code and descriptive text in a single document, an increasingly large community of researchers have adopted computational notebooks to document, publish and share their research via personal websites and GitHub (Parente 2019).

Yet publishers have not embraced this technology. We believe that there are great benefits for the scientific community and general public from publishing computational notebooks. The publication of computational notebooks, along with articles, enables reproducibility and replicability of data analysis and methods. Computational notebooks offer a valuable vehicle for teaching and demonstration of analytical tools. They can also augment the impact of research beyond its primary objectives by extending original analysis and by reaching non-academic communities (Arribas-Bel et al. 2020). The interactivity of notebooks can engage policy makers and the general public in ways that standard academic journal publications cannot. Notebooks can be used to engage policy, discipline-specific or local knowledge experts in the research process. In doing so, data and outcomes channeled through notebooks can enable the identification of new relevant patterns or uses that may have not been reported or explicitly discussed in the original publication.

In view of these potentials, REGION officially announced a new form of publication, computational notebooks, in 2019. In order to demonstrate the value of computational notebooks in regional research and to stimulate this means of publication, we organized this special issue. REGION will continue to accept submissions in computational notebooks (.ipynb and .Rmd files). When accepted, computational notebooks are published in three formats: R or Python notebook file extensions, HTML and pdf. Unlike the pdf format, R or Python notebook file and HTML file extensions will provide an interactive version of the code which can be fully reproduced. REGION encourages authors to make submissions in these formats. Two publication options are available publishing computation notebooks: (1) as a companion to a research article, or (2) as standalone piece in the Resource section. By publishing computational notebooks, REGION seeks to encourage appropriate recognition of the work dedicated to this form of document. Normally the use of computer code published on personal websites or GitHub does not receive appropriate recognition by the way of citation given unfamiliarity with this form of publication or lack of a referenceable identifier. In REGION, computational notebooks are published as regular papers, receive a digital object identifier (DOI), and hence will be referenceable and citable.

2 The Issue

This Special Issue aims to introduce the publication of computational notebooks in REGION. Seven articles make up this Special Issue and illustrate some of the the key benefits of computational notebooks.

The first three articles use notebooks to introduce advanced urban planning and statistical methods available through open software. Boeing (2019) illustrates the potential of computational notebooks in urban analytics and planning introducing 'OSMnx'. 'OSMnx' is a Python package for working with OpenStreetMap data and modelling, analysing and visualising street networks anywhere in the world. The notebook shows how to download and model street networks, compute network indicators, visualise street centrality, calculate routes, and work with other spatial data, including building footprints and points of interest.

Sarrias (2020) uses a computational notebook to introduce a R package called 'Rchoice'. Rchoice offers a statistical modelling framework to estimate spatial heterogeneity by estimating locally varying coefficients offering different latent structures in a discrete choice setting. Wieland (2019) introduces the R package 'REAT', a Regional Economic Analysis Toolbox for R, and extensively illustrates its capabilities with regional economic data for Germany. Together the computational notebooks by Boeing, Sarrias, and Wieland illustrate how methods can be introduced to new users and help researchers reach broader audiences interested in learning from, adapting, and remixing their work.

The following two articles use computational notebooks to introduce the application of novel methods and data. Comber (2019) illustrates the use of machine learning to conduct address matching. Data often lack of unique identifiers to enable one-to-one address matching. Deterministic matching based hand-crafted rules that classify address matches and non-matches based on specialist domain knowledge are typically applied. Machine learning approaches can provide a faster and automatable way to match addresses with little human intervention. The notebook offers an end-to-end pipeline to conduct address match using machine learning.

Chen et al. (2020) contributes a computational notebook to acquire, process and analyse satellite imagery. While satellite imagery is often used to study and monitor changes in natural environments and the Earth surface, it has remained underutilised in Regional Science to study cities despite the open availability and extensive temporal coverage of data sets, like Landsat enabling monitoring long-term changes for a period of up to 46 years. The notebook offers a tool to demonstrate how to batch-download high-resolution satellite imagery; and enable the extraction, analysis and visualisation of features of the built environment to capture long-term urban changes.

Patias (2019) contributes a notebook illustrating its interactivity potential as an engaging resource for end users and researchers through a regional analysis of youth unemployment in Europe. Interactive maps and figures enable readers to explore the data in greater detail by dragging, brushing and zooming. The notebook also offers an end-to-end workflow from reading raw data via API, through the data processing, to the final publication outputs. The notebook demonstrates the existence of systematic pattern of youth unemployment across Europe. Four distinctive groups of regions are identified: 'stable low youth unemployment'; 'stable moderate youth unemployment', 'increasingly high youth unemployment', and 'stable high youth unemployment'.

The final article by Reades (2020) illustrates the use of computational notebooks to support teaching delivery and enhance student learning. Specifically, the notebook argues that given the proliferation of large and complex spatial data, there is a need not only for quantitative skills, but also for computational skills. The notebook also shows how computational notebooks can assist in developing and delivering a suite of geo-computational modules to enhance data science and analytics skills.

Together, the articles in the Special Issue demonstrate how notebooks can be used to introduce the operation of open software and application of novel methods, enhance student learning, increase interactivity and exploration of research outputs, and how to produce replicable, reproducible and transparent research. We encourage submissions to this new form of publication. We believe that computational notebooks offer an exciting new platform adhering to REGION's principles of open, reproducible and transparent science, and anticipate a change in the future of academic publishing in this direction.

Acknowledgments

Sergio J. Rey acknowledges the support of NSF-SES 1831615.

References

Arribas-Bel D, Green M, Rowe F, Singleton A (2020) Open data products – A framework for creating valuable analysis ready data. *Journal of Geographical Systems*. in review

Begley CG, Ellis LM (2012) Drug development: Raise standards for preclinical cancer research. *Nature* 483: 531–533. CrossRef.

Begley S (2012) In cancer science, many "discoveries" don't hold up. Reuters. http://www.reuters.com/article/us-science-cancer-idUSBRE82R12P20120328

Boeing G (2019) Urban street network analysis in a computational notebook. *REGION* 6: 39–51. CrossRef.

Branco A (2012) Reliability and meta-reliability of language resources: Ready to initiate the integrity debate? The 12th workshop on treebanks and linguistic theories (tlt12)

Branco A, Bretonnel Cohen K, Vossen P, Ide N, Calzolari N (2017) Replicability and reproducibility of research results for human language technology: Introducing an LRE special section. *Language Resources & Evaluation* 51: 1–5. CrossRef.

Chen M, Fahrner D, Arribas-Bel D, Rowe F (2020) A reproducible notebook to acquire, process and analyse satellite imagery: Exploring long-term urban changes. *REGION* 7: 15–46. CrossRef.

Comber S (2019) Demonstrating the utility of machine learning innovations in address matching to spatial socio-economic applications. *REGION* 6: 17–37. CrossRef.

Editorial (2013a) Announcement: Reducing our irreproducibility. Nature News Nature

Editorial (2013b) Unreliable research: Trouble at the lab. The Economist. http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

Fanelli D (2009) How many scientists fabricate and falsify research? A systematic review and metaanalysis of survey data. *PloS ONE* 4: e5738. CrossRef.

Fokkens A, van Erp M, Postma M, Pedersen T, Vossen P, Freire N (2013) Offspring from reproduction problems: What replication failure teaches us. *Proceedings of the 51st annual meeting of the association for computational linguistics* 1: 1691–1701

Ioannidis JP (2011) An epidemic of false claims. *Scientific American* 304: 16–16. CrossRef.

Jupyter Project (2019) Jupyer notebooks. Available at: https://jupyter.org (accessed: 1 September 2019)

Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C, Jupyter Development Team (2016) Jupyter notebooks – A publishing format for reproducible computational workflows. In: Loizides F, Schmidt B (eds), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS-Press, Amsterdam, The Netherlands, 87–90. CrossRef.

Koster S, Rowe F (2019) Fueling research transparency: Computational notebooks and the discussion section. *REGION* 6: 1–2. CrossRef.

Naik G (2011) Scientists' elusive goal: Reproducing study results. Wall Street Journal, December 2 2011, a1

Open Science Collaboration (2015) PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 349: 943–950. CrossRef.

Parente P (2019) Estimate of public Jupyter notebooks on GitHub. Github, available at: https://github.com/parente/nbestimate (accessed: 5 September 2019)

Patias N (2019) Exploring long-term youth unemployment in Europe using sequence analysis: A reproducible notebook approach. *REGION* 6: 53–69. CrossRef.

Peng R (2011) Reproducible research in computational science. *Science* 334: 1226–1228. CrossRef.

Pérez F, Granger B (2015) Computational narratives as the engine of collaborative data science. Blog. available at: https://blog.jupyter.org/project-jupyter-computationalnarratives-as-the-engine-of-collaborative-data-science-2b5fb94c3c58 (accessed: 10 September 2019)

Prinz F, Schlange T, Asadullah K (2011) Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 10: 712–712. CrossRef.

Reades J (2020) Teaching on Jupyter – Using notebooks to accelerate learning and curriculum development. *REGION* 7: 21–34. CrossRef.

Sarrias M (2020) Random parameters and spatial heterogeneity using Rchoice in R. *REGION* 7: 1–19. CrossRef.

Wieland T (2019) REAT: A regional economic analysis toolbox for R. *REGION* 6: R1–R57. CrossRef.

Zimmer C (2012) A sharp rise in retractions prompts calls for reform. The New York Times 16. http://www.nytimes.com/2012/04/17/science/rise-in-scientific-journal-retractions-prompts-calls-forreform.html