Levi John Wolf
School of Geographical Sciences

ORCID: https://orcid.org/0000-0003-0274-599X

Geographic Data Science for All

submitted for the category of

Widening Reach

in the

Open Research Prize 2022

University of BRISTOL

SUMMARY (100 word abstract)

Across my professional life, I have comprehensively adopted open science practices. This includes pre-printing papers, publishing replication materials, releasing "production-ready" software (with tests and documentation) for all publications, actively finding and fixing reproduction/replication failures, mentoring and training diverse early career researchers to join open scientific software communities, and leading large international bids to obtain funding for community-owned scientific infrastructure. Overall, this enriches and sustains core infrastructure in social and environmental sciences and builds community. Open science is integral to how I work, collaborate, teach, and mentor.

What did you do?

I adopt open practices across every level of my science. Every paper I lead is pre-printed (alongside replication data and code) in the Open Science Framework. Upon publication, these papers[i] ground contributions to packages in the Python Spatial Analytics Library (PySAL), a collection of 15 software packages supporting spatial analysis in Python[ii], (19k downloads/month), or geopandas (2 million downloads/month), a core Python[ii] library for geographic data analysis. Software I (co)author and maintain is used in industry,[iii] third sector,[iv] and both within[v] and outside[vi] my academic discipline. I actively identify and contribute fixes to reproducibility failures.[vii] In the last year, I have led two major international funding bids (>£600k) for infrastructure maintenance and development. On the pedagogical side, I have also mentored twelve Google Summer of Code (GSoC) students over five years (including a University of Bristol undergraduate) with projects to improve geospatial software and/or provide public implementations of papers in my field. Finally, I am writing a textbook (with co-authors) on GitHub that is currently available online under a CC-BY-NC-ND license and in hardback by CRC Press in 2023. Thus, across research and teaching, I adopt open science practices.

Why did you do it?

Community-owned scientific infrastructure is critical for open science; the practices discussed above contribute to this. Historically, proprietary software from companies that commercialized cutting-edge research on geographic information systems supported novel social and environmental science. Over time, though, this "walled garden" has insulated practitioners from more fundamental issues around algorithms, data, and representation, limiting scientific progress (Gahegan, 1999; Wolf et al. 2021). In contrast, geographic data science (Singleton & Arribas-Bel, 2019) represents our entrance into a new era: community-owned scientific infrastructure that I help build, maintain, and grow is now empowering a new generation of scientists to do novel and innovative work on their own terms.

What barriers / challenges did you have to overcome?

I have dealt with three distinct challenges in my experience building and maintaining scientific infrastructure. First, it can be scary to start contributing to public scientific software. When you publish code for others to use, any bug report can feel like an attack on your work or feature request an imposition on your time. Further, publishing your scientific code can invite very detailed critique. To address this challenge, it helps to try to stay grounded about what

you build; critique code & concepts, not people; and seek outside advice/perspective when you feel insecure. Second, while writing *new* software may be research-adjacent, *maintenance* is rarely incentivised in academia. This is true for open software generally (Eghbal, 2016): credit and recognition is rarely proportionate to use. Structural lack of incentives can eventually sap motivation, leading to a burnt out and moribund developer community. Thus, I try to invest personally in my fellow developers, take occasional breaks from maintenance, and use my journal editorship to create a better incentive structure for the next generation (Arribas-Bel et al. 2021). Finally, it can be hard to build and maintain a diverse community for scientific software. Therefore, equity and diversity in community-building is critical for our future. So, for example, GSoC students I seek to mentor are predominantly female or from the Global South.

What does it mean for you and your research?

These practices are how I do science; they are not "extras" to some normal practice. Thus, my fellow developers are my co-Investigators on grants and co-authors on papers. We train the next generation in these practices together. My science would be impossible without this community, and I would not have entered or stayed in academia without it.

How might your findings / approach help other researchers?

Using code as text (Rey, 2009) can help bring abstract statistical or conceptual approaches "down to Earth," as sequences of simple steps build to something larger. Sharing this code builds a teaching community, too (Arribas-Bel 2019; Reades & Rey 2021). Open scientific infrastructure can also be liberatory: high-quality free and open alternatives to expensive proprietary systems can be improved, redistributed, and verified by users. Finally, investing in scientific infrastructure can lead to cross-pollination between domains. At its best, the epistemic openness fostered by community-owned scientific software can break down academic knowledge silos and build new interdisciplinary collaborations from shared infrastructure.[v] Thus, investment in community-owned scientific infrastructure benefits all.

Additional Information

[i] For example, novel segregation statistics in Wolf et al. (2019) or replication failure resolutions by Sauer et al. (2021) are implemented in the esda package, and spatial multilevel model methods from Wolf et al. (2021) are in the spvcm package. These implementations are designed for end-users and (as such) go through user experience & design review, are unit-tested, and have extensive documentation. They are also available *in addition to* replication materials.
[ii] Python is highly ranked (TIOBE, PYPL, StackOverflow) and used in industrial & academic data science (Donoho, 2017).
[iii] This demo of Microsoft's Planetary Computer uses geopandas to visualize Hurricane Florence, CARTO provides part of PySAL (e.g. Oshan et al. (2021)) in their webapp, and 48% of our 2020 User Survey (~96 people) are commercial.
[iv] We support users and developers at the Ordnance Survey, the European Space Agency, are an "important package" for analysts to learn by Eurostat's training agency, and 30% of our 2020 User Survey (~60 people) are 3rd sector users.
[v] The package's original announcement (Rey & Anselin, 2007) has been cited 87 times on Scopus, and we have directly confirmed 37 of these uses, and 1000+ papers mentioning the package or citing it incorrectly in Google Scholar.
[vi] A recent example is Palla et al. (2022)'s new "omics" microbiology library which borrows heavily from PySAL, referring to our implementations often in their code and, in cases, integrating our algorithms and estimators directly.
[vii] In addition to production-ready Python implementations mentioned above, improvements from Sauer et al. (2021) were also contributed to the R package spdep (38k monthly downloads) by Sauer (a GSoC student) and me.