Reproducing and Replicating Spatial Data Science

Joseph Holler, Peter Kedron, and Sarah Bardin

HEGSRR.github.io

Funding Support from NSF BCS-2049837, NSF OAC-1743184

Middlebury UCSB Arizona State University

Reproduction and Replication

Framing expectations

Workshop Agenda

Time	Topic	Presenters
15 min	What is R&R and why should you care?	Kedron, Bardin
25 min	Which open science practices are you familair with?	Holler, Bardin
15 min	What else do you want to know 1? (Discussion)	All
5 min	Break	All
10 min	What does the open science ecosystem look like?	All
30 min	What can open geographical science look like in practice?	Holler
5 min	How do you get started? Where can you contribute?	Kedron, Holler
15 min	What else do you want to know 2? (Discussion)	All

Central Objective:

1) Situate R&R in the geography and spatial data science

&

2) introduce selected research practices and resources you can use to make your work more reproducible

Tell us what you think

Please take 2mins to complete a short survey about open science research practices

Protocol

- Sort 11 open and reproducible research practices (ORRP)
  - Already using
  - Aware and interested in using
  - Unaware or uninterested
- 10 likert technology adoption questions
- Share deidentified data
- Follow-up surveys (immediate, one year)

Defining Reproducibility and Replicability

A brief definition for spatial data science

Reproduction and Replication

(Schmidt 2009, Gomez et al. 2010, Barba 2017, Christensen et al. 2019, NASEM 2019)

TABLE 1. Types and Purpose of Replication

Reproduced from Christensen et al. (2019, p159, Table 9.1)

Compared to original study	**Focused on *repeating procedures***	**Focused on *introducing differences***
***Same* data**	Verification	Reanalysis
***Different* data**	Direct Replication	Extension

Reproduction and Replication

(Schmidt 2009, Gomez et al. 2010, Barba 2017, Christensen et al. 2019, NASEM 2019)

TABLE 1. Types and Purpose of Replication

Reproduced from Christensen et al. (2019, p159, Table 9.1)

Compared to original study	Focused on *repeating procedures*	Focused on *introducing differences*
***Same* data**	Verification	Reanalysis
***Different* data**	Direct Replication	Extension

Reproduction

Same data, same procedure, same results, same context

Reproduction and Replication

(Schmidt 2009, Gomez et al. 2010, Barba 2017, Christensen et al. 2019, NASEM 2019)

TABLE 1. Types and Purpose of Replication

Reproduced from Christensen et al. (2019, p159, Table 9.1)

Compared to original study	Focused on *repeating procedures*	Focused on *introducing differences*
***Same* data**	Verification	Reanalysis
***Different* data**	Direct Replication	Extension

Reproduction

Same data, same procedure, same results, same context

Replication

New data, similar procedure, similar results, same or new context

Veridical Spatial Data Science

Munafo et al. (2016), Kedron et al. (2020), Yu & Kumbier (2020)

Principled inquiry to extract reliable and reproducible information from spatialtemporal data, with an *enriched technical language* to communicate and evaluate empirical evidence in the context of human decisions, domain knowledge, and geographic confounds; *supported by a system of external validation and evidence accumulation based on the purposeful replication of findings across space and time*.

(Adapted from Kedron and Bardin 2021, Yu and Kumbier 2020)

Missing Replication Studies

Geographers are thinking about R&R, but not attempting replications

BCS-2049837

Reproducibility

81% Consider reproducibility of their own work

71% Discuss reproducibility with colleagues

53% Consider reproducibility during peer-review

Replicability

74% Consider replicability of their own work

65% Discuss replicability with colleagues

59% Consider replicability during peer-review

14%

Reported attempting a reproduction
(**7%** attempt publishing)

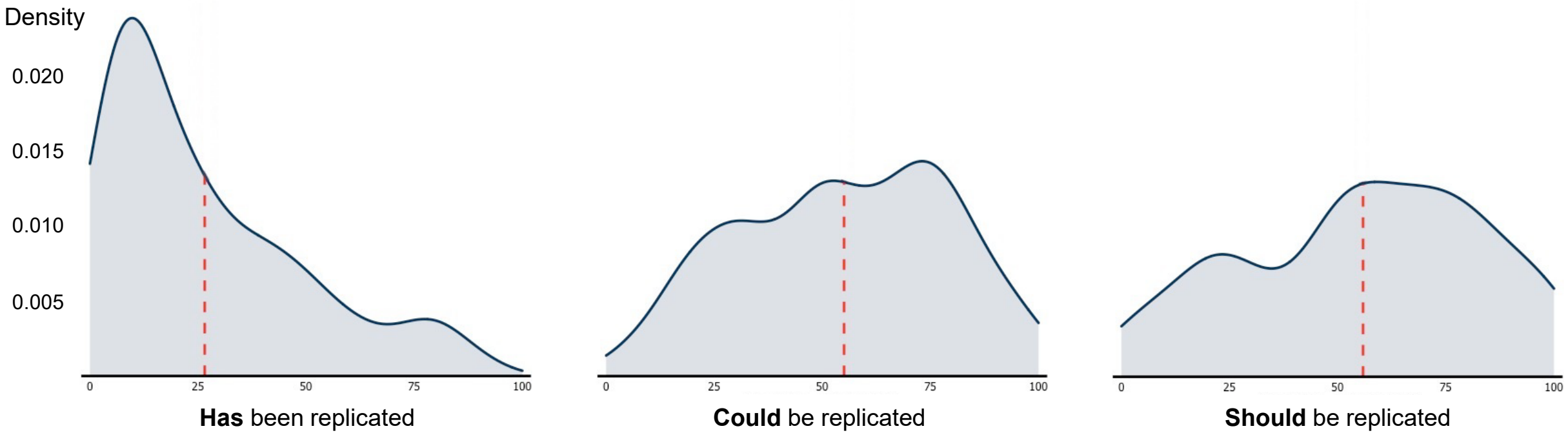
31%

Reported attempting a replication
(**21%** attempt publishing)

Missing Replication Studies

What percentage of geographic research do you believe …

BCS-2049837



2023:
The Year of Open Science

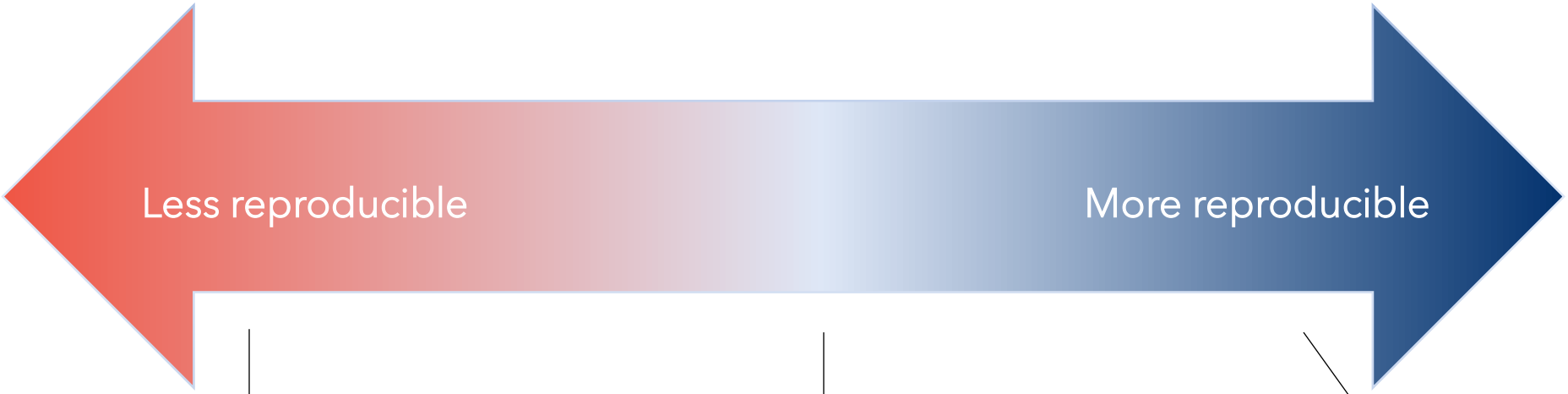
Practicality of R&R in the Policy Environment

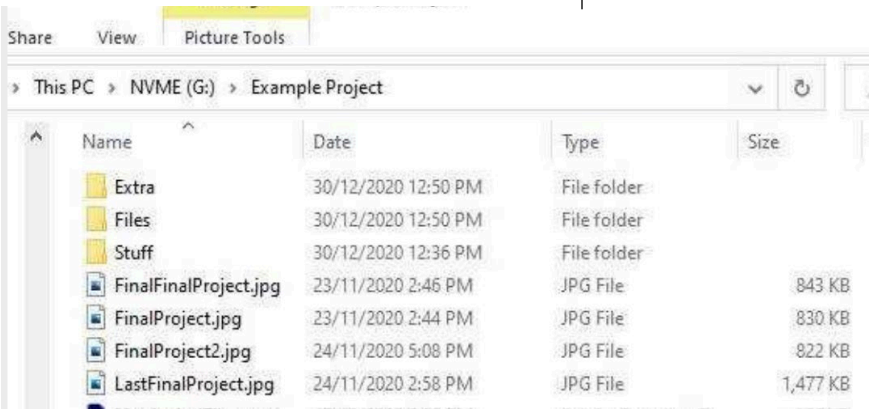
Reproducible workflows are the pragmatic solution to large-scale research studies

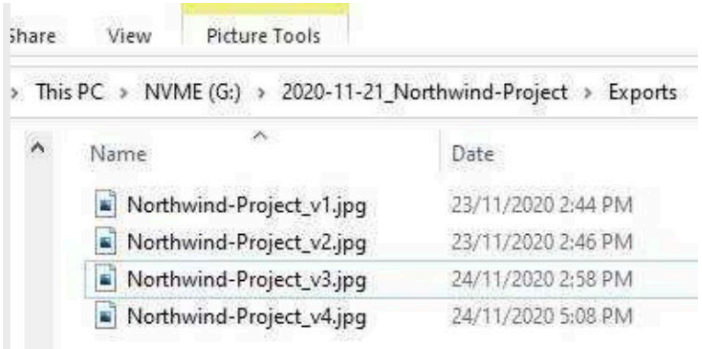
- Provides **quality assurance** that the numbers published in reports are **accurate and error-free**
- **Improves the efficiency** of conducting large-scale program and policy evaluation
- **Enhances transparency** of research allowing for **scrutinizing of assumptions, data sources, and methods**

R&R as a Matter of Practice

It is all a spectrum







Our R&R Related Resources

Munafo et al. (2016), Kedron et al. (2020), Yu & Kumbier (2020)

hegsrr.github.io

hegsrr.github.io/Workshop-SDSS-2023/

- 5 Peer-reviewed Publications
- 8 Reproduction and Replication Studies
- 2 Surveys of Researcher Practices
- **Reproducible Project Repository Template**
- Manual In Development
- 2 Course Syllabi
- 9 RAs Mentored
- ~50 Students Engaged in R&R Studies

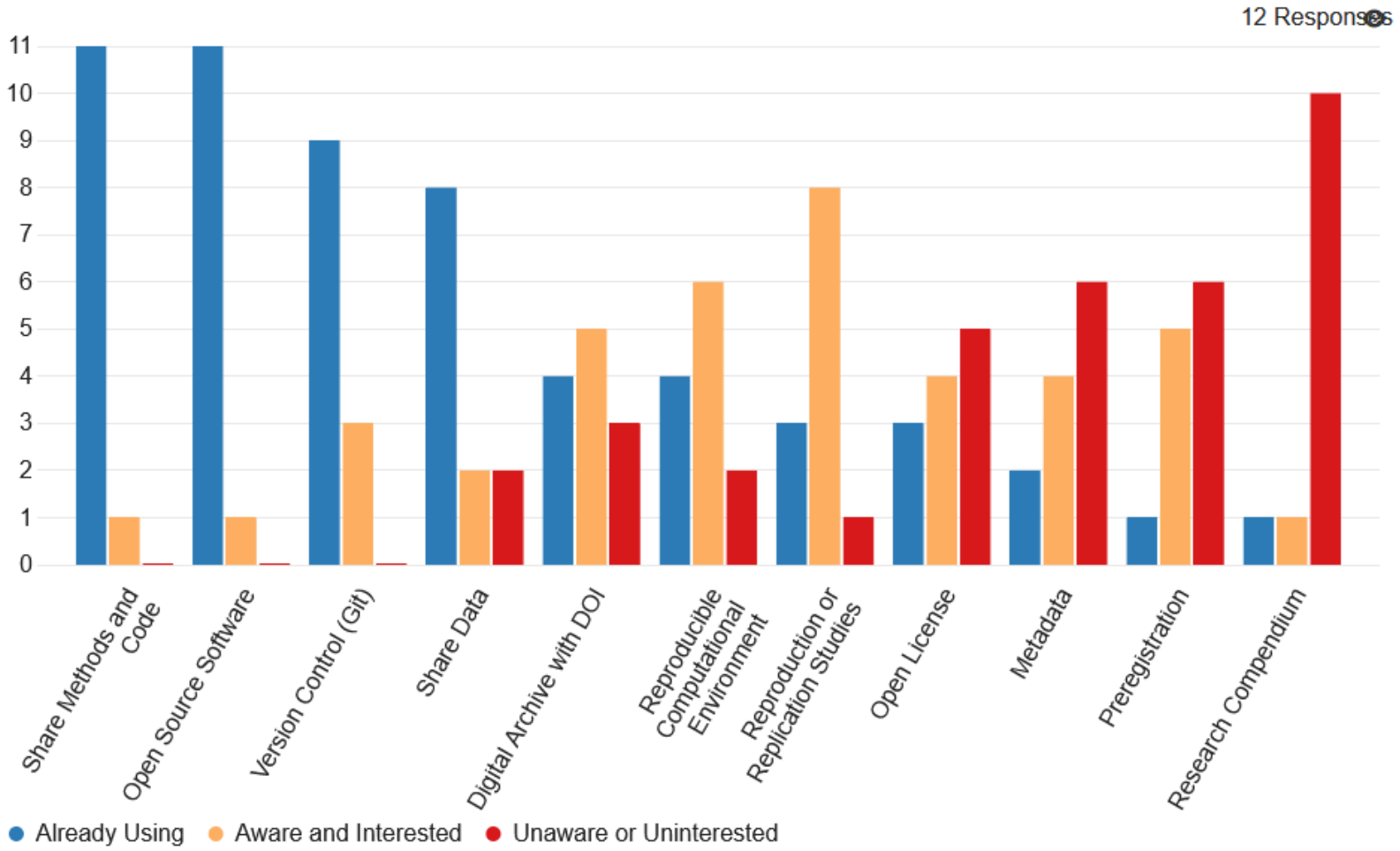
Open and Reproducible Research Practices

A quick review of practices in light of the survey results

Survey Results: You

Goal: incremental progress as individuals and scientific community

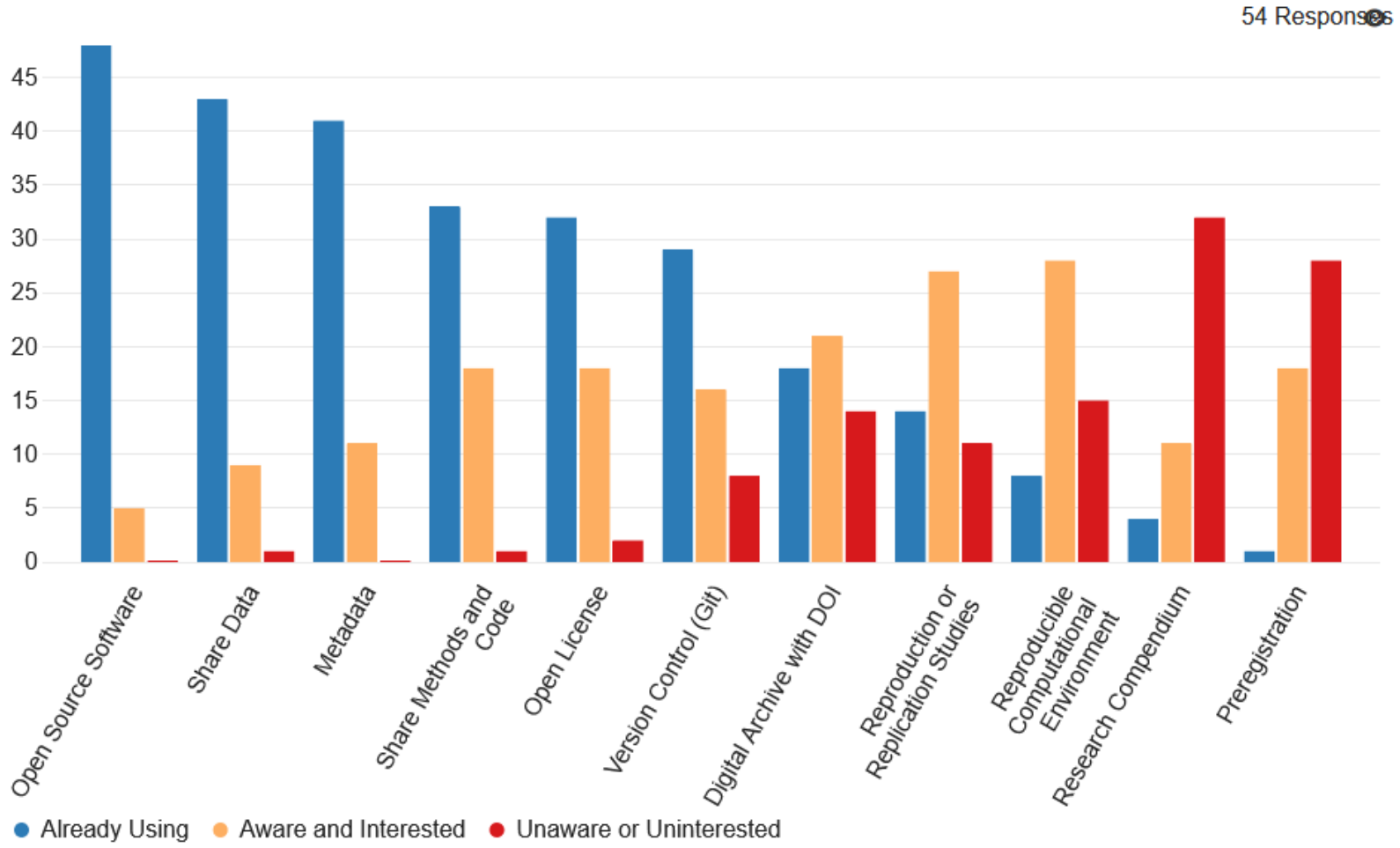
Be kind to your future self!



Results: Spatial Data Science Symposium

Goal: incremental progress as individuals and scientific community

Be kind to your future self!



Share Methods and Code

Rmarkdown, Jupyter Notebook, Scripts, Models, Protocols

Do you share a complete description of your methods?

Open Source Software

Python, R, PostGIS, GeoDA, QGIS

Do you use and cite research software with (re)distributable source code?

Version Control

Git, GitHub, GitLabs, OSF projects

Do you manage and track changes
to your study design, data, and code?

- Best at tracking one line of plain text
- GitHub integrates with Overleaf, OSF, Webpages

Share Data

Do you make the data for your study readily available in the most complete and unmodified form permissible by law and ethical protocols?

Digital Archive with DOI

DOI: Digital Object Identifier

Are all the components of your study digitally archived for long-term preservation, and labelled and linked with a persistent digital object identifier?

Reproducible Computational Environment

Docker container, public cyberinfrastructure, environment metadata

Do you provide access to your computational environment or sufficient information about your environment such that others can recreate it?

Reproduction and Replication Studies

Do you attempt and share reproduction or replication studies?

Open Licensing

Creative commons, BSD-3, GPU, MIT…

Do you license your research products to allow others to use, modify, and redistribute them?

Metadata

Dublin Core, ISO 191**, Federal Geographic Data Committee (FGDC)

Do you provide information about your study and each of its components in a standardized format?

Preregistration

OSF, AsPredicted, Registered Reports

Do you publicly register your hypotheses and research design before conducting your work?

Research Compendium

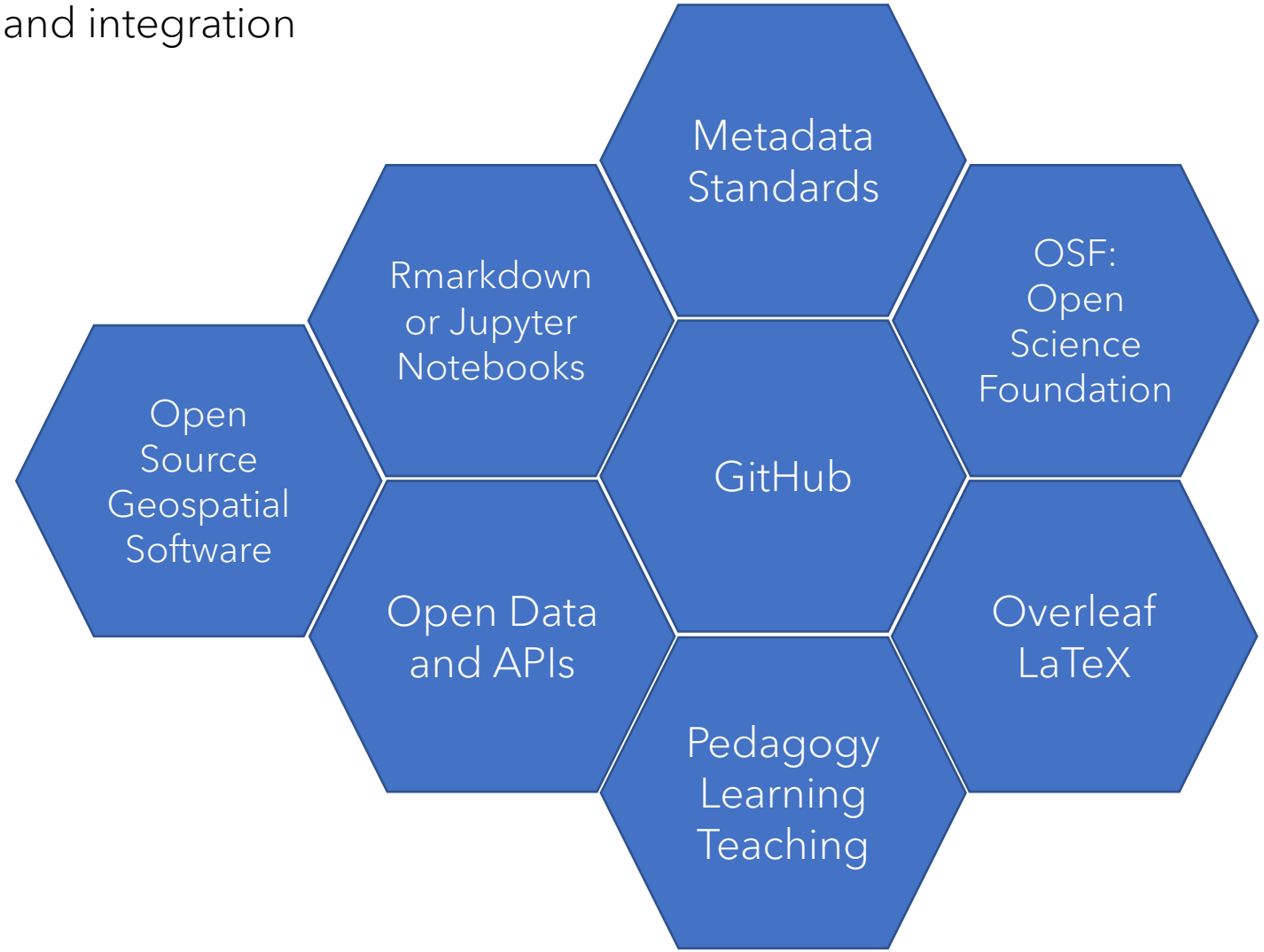
HEGSRR Template, TIER Protocol, WORCS, o2r

Do you collect *all* components of your study together in a directory organized with consistent structure and relative links?

Open Science Ecosystem

Multiple points of entry and integration

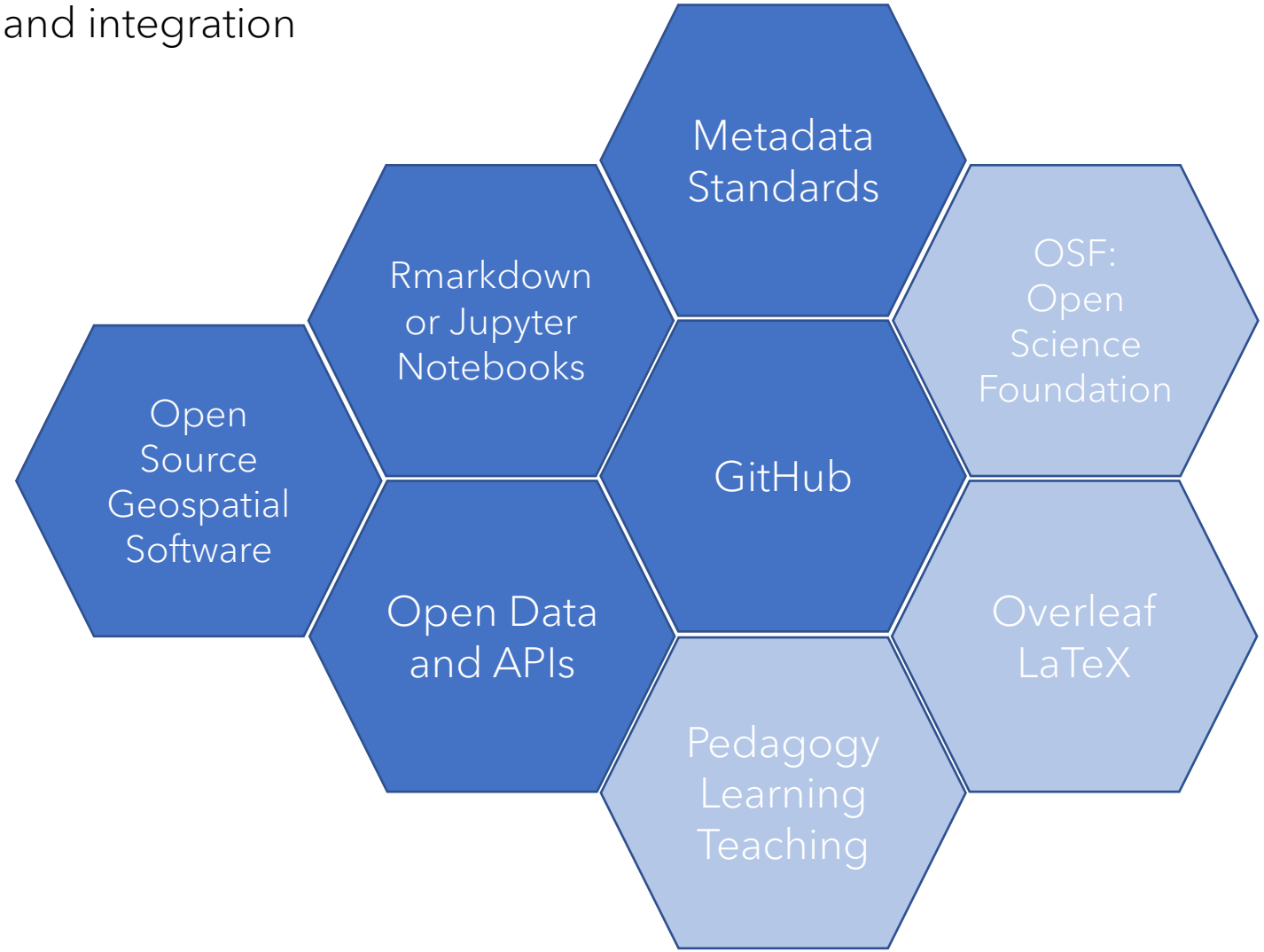
Where are you in your own research project and training?



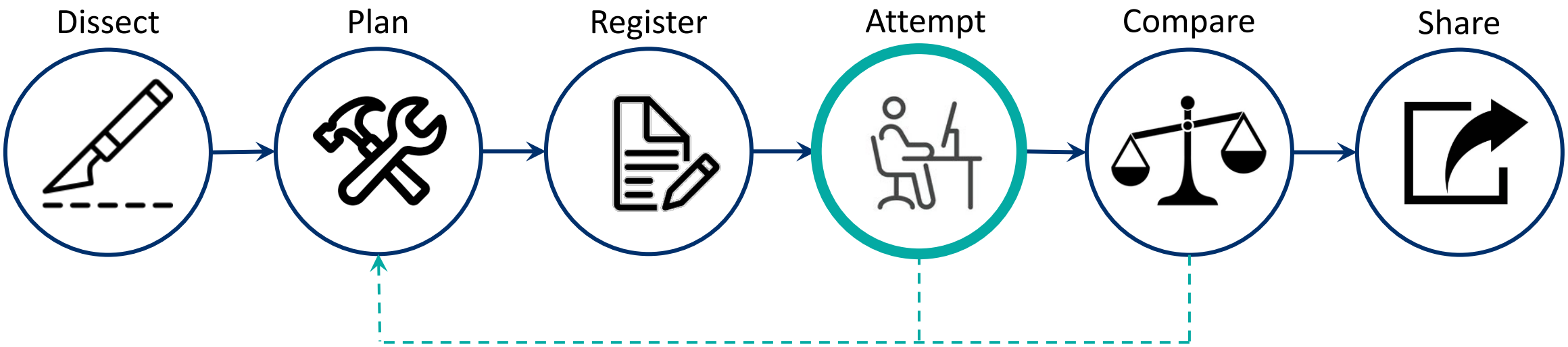
Open Science Ecosystem

Multiple points of entry and integration

In one undergraduate course…



Conducting Replications of Geographic Research



OSF

Pre-Analysis Template
Registration Guide

GitHub

Repository Templates
Version Control
Issue Forum

OSF

Pedagogical Outline, PAP templates, and Repository Structures are all available Reproducible, Replicable, and Open Science Practices in the Geographical Sciences site (https://osf.io/c5a2r/) and HEGSRR Github Organization (https://github.com/HEGSRR)

Some Typical Outputs

Open access data, methods, reports for research and teaching

[RP] Report

HE&G ReScience

Reproduction of

Beyond the 405 and the 5: Geographic Variations and Factors Associated With Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Positivity Rates in Los Angeles County

by Vijayan T., Shin M., Adamson P.C., Harris C., Seeman T., Norris K.C., Goodman-Meza D.

Created 18 February 2021

Revised 05 July 2021

Reproduction Authors:

Peter Kedron[1,2] / Joseph Holler[3] / Sarah Bardin[†,1,2] / Summer Cliff / Kimberly Fuller / Joshua Gilman / Bryant Grady / Megan Seeley / Addison Van Zanbergen / Wenxin Wang / Xin Wang

[1]Arizona State University, School of Geographical Sciences and Urban Planning, Tempe, AZ, USA
[2]Arizona State University, Spatial Analysis Research Center (SPARC), Tempe, AZ, USA
[3]Middlebury College, Department of Geography, Middlebury, VT, USA
[†]Graduate Student Author

Replication Materials Available at:
Pre-registered Plan – https://github.com/HEGSRR/RP-Vijayan-2020/tree/main/docs/report
Data – https://github.com/HEGSRR/RP-Vijayan-2020/tree/main/data/private
Code – https://github.com/HEGSRR/RP-Vijayan-2020/tree/main/procedure/code

Correspondence should be addressed to Peter.Kedron@asu.edu. The authors declare that no competing interests exist

Research Hypotheses to Reproduce

H1: There is a difference in mean values of key socioeconomic and demographic variables by positivity rate groupings of low, medium, and high areas.

Original test: One-way analysis of variance (ANOVA) indicated that all variables, except the percentage black, exhibited statistically significant differences among the three subgroups (Table 1).

H2: (a) COVID-19 age-adjusted testing rate, (b) age-adjusted diagnosis rate, and (c) crude positivity rate were non-randomly distributed throughout LA County.

Original test: Local indicators of spatial association (LISA) identified elevated values of each variable around the center of LA, and depressed values around the edges of the county (Fig. 1)

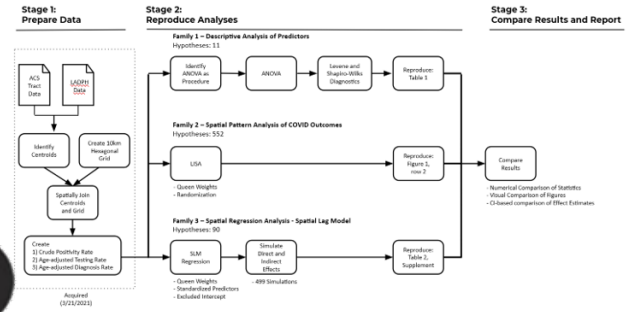
H3: Socio-structural characteristics of LAC have non-zero association with crude positivity rate.

Original test: The authors used a regression model with a spatially lagged response was to identify significant positive associations between crude positivity and (i) proportion of population over 65, (ii) proportion Latino, proportion living in poverty, and (iii) housing density.

Key Findings

- We were able to reproduce the original analyses after contacting the authors to obtain their data file. Original effect estimates for the regression coefficients fell within the 95% confidence intervals of our reproduction.
- The lack of details concerning how the hexagons that were used as the unit of analysis by the authors were created, or how data from census tracts were aggregated to those hexagonal units inhibited our ability to independently reproduce the original results. Moreover, these same issues made it difficult to assess what the predictor and response variables were in fact measuring as this spatial aggregation potentially misaligned several variables with the conventional definitions.
- The authors did not provide equations or formulations related to their implementation of the SLM analyses, it is therefore difficult to assess how the models should be properly interpreted. The language used in the original discussion by the authors imprecisely describes the coefficients reported and ignores the fact that these are based on standardized variables and that the model intercept was omitted from the analysis. These issues raise questions about how the estimated effects should be interpreted.

HE&G ReScience — Kedron et al. (2021) 1



1 Reproduction workflow
ashed lines indicate steps that were completed by original author and supplied upon request for purposes of the on.

d Differences from the Original Study: Vijayan et al. (2020) did not specify the computational onment or software used during their original study. In the absence of this information, we decided to lement these analyses in R v4.0.5 and Rstudio v1.4.1106 using the following packages: tidyverse, sp, al, spatialreg, spdep. Although Vijayan et al. described using a permutation approach for identifying tatistically significant clusters in their LISA analyses, they did not detail the number of simulations conducted, nor did they indicate how statistical significance was calculated for the SLM models. We elected to implement a permutation approach which used 499 simulations in order to calculate our p-values. Because there is inherent randomness in the permutation approach, when a seed is not set, and we cannot be certain of the number of simulations performed, we did not expect to be able to fully reproduce the exact p-values reported in the paper, however we expected that the direction and magnitude of the results would be consistent between the original analysis and the reproductions.

Assessment Criteria: As noted earlier, we did not anticipate being able to achieve bitwise reproduction of any of the LISA or spatial regression results with those from the original study. As a result, we compared the direction, magnitude, and significance of our results with those of the original authors.

Reproduction Results

Reproduction Results for the Descriptive Statistical Analyses of Predictors (H1):

The first set of hypotheses examined by Vijayan et al. compared the distribution of a set of predictor variables across three subgroups of low, medium, and high positivity rates. Vijayan et al. did not specify the type of correlational analysis performed, so we performed one-way ANOVA tests. As shown in Table 1 below, we achieved bitwise reproduction of the original analysis results. That is, all means, standard deviations, and reported p-values from the original study were identically reproduced.

HE&G ReScience — Kedron et al. (2021) 4

Kedron et al. (Revision), *Geographical Analysis*

HEGS-RR Infrastructure

A project-based demonstration of our infrastructure for reproducibility and replicability

Questions about Infrastructure?

We invite questions, feedback, and discussion about R&R in spatial data science

Open Discussion

We invite questions, feedback, and discussion about R&R in spatial data science

Discussion prompts

and invitation to collaborate…

1. Can you share any *successes*, *advice*, or *best practices* introducing reproducibility and replicability in your own scholarship (research or teaching)?

2. What *barriers* do you perceive to adopting open and reproducible research practices in your own scholarship?

3. Could any *resources, changes,* or *incentives* help overcome those barriers?

Thank You

Our R&R Related Resources

hegsrr.github.io

/Workshop-UCSB-2023

- 5 Peer-reviewed Publications
- 8 Reproduction and Replication Studies
- 2 Surveys of Researcher Practices
- **Reproducible Project Repository Template**
- Manual In Development
- Course Syllabi
- 9 RAs Mentored
- ~50 Students Engaged in R&R Studies