Teaching Reproducibility and Replicability in Spatial Data Science

Where to Start and What to Do

Peter Kedron University of California Santa Barbara
Joseph Holler Middlebury College
Andrew Trgovac Arizona State University

HEGSRR.github.io

Funding Support from NSF BCS-2049837

Dr. Peter Kedron

Associate Professor
Department of Geography
UC Santa Barbara

Associate Director
Center for Spatial Studies & Data Science

Dr. Joseph Holler

Associate Professor
Department of Geography
Middlebury College

Dr. Andrew Trgovac

Teaching Assistant Professor
School of Geographical Science and Urban Planning
Arizona State University

Workshop Agenda

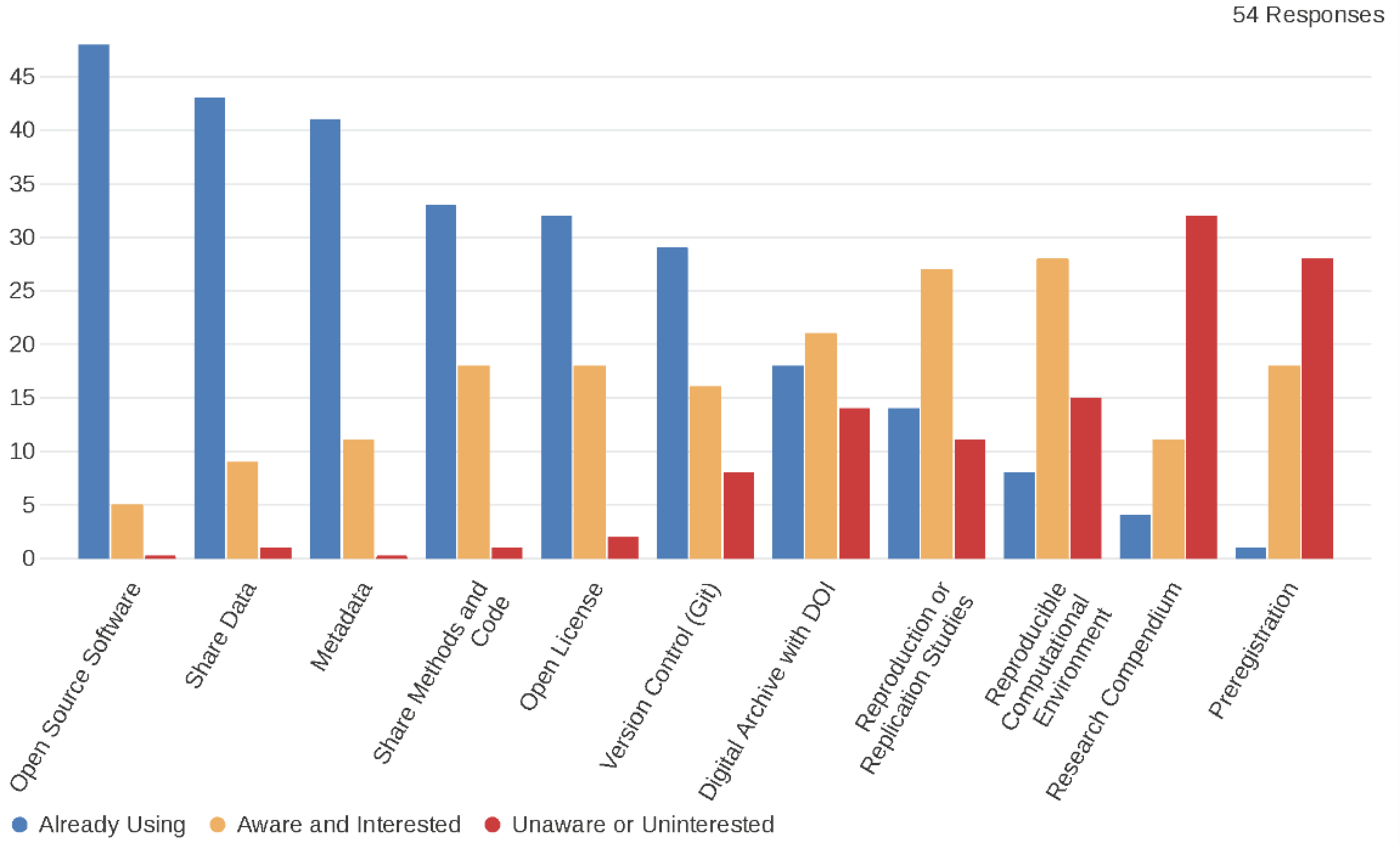
Introduce a competency rubric and a classroom model for R&R education in spatial data science

Table 1. Workshop Agenda and Presenters

Time	Topic
10 min	Introduction and Major Takeaways
10 min	Competency Rubric
10 min	Q&A
05 min	Classroom Model
05 min	Project Based Learning
20 min	Implementing the Classroom Model
10 min	Q&A
05 min	Resources and Next Steps

Open Science Practices

More than sharing data and code



A Working Definition

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

	Purpose	Data	Context	Procedure	Results
Reproduction	Internal Validity	Same	Same	Same	Same
Reanalysis	Internal Validity	Same	Same	Different	Similar
Replication	External Validity	Different	—	Similar	Similar

A Working Definition

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

Veridical Spatial Data Science

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)

Key Ideas for Teaching Reproducibility and Replicability

What to take away from this workshop

(1) Teach more than data and code sharing.
Link R&R to the epistemology of science and open science

(2) Attempt reproductions and replications with your students
Pedagogically rich way to teach GIScience, spatial data science, and topical knowledge

(3) Reproduction attempts create intrinsic and extrinsic rewards
Reproduction attempts are rewarding for students, improve learning outcomes, produce publications

(4) Use (and improve) our open educational materials
Templates, past reproductions, teaching materials

Reproducibility in Spatial Science

Competencies and Metrics for Curriculum Development and Learner Assessment

Inspiration For This Work

Competencies of reproducible spatial data science

2024

O'Donnell, K. L., Aiello-Lammens, M., Bledsoe, E., Bowlick, F. J., Broughton, L., Calderon, O., Crispo, E., Emery, N., Farrell, K., Ngiramahoro, M., Patel, N., Paudel, S., Richardson, L., Soares, B. E., Supp, S., Weigel, E.

BEDE Network Data Science Skills Curriculum Map. Biological and Environmental Data Education (BEDE) Network

QUBES Educational Resources

https://qubeshub.org/publications/4859/alignments/1

Structure of the Reproducibility Competency Rubric

We identify four competency categories and three levels of achievement

Category	Skill	Foundational	Second Level	Third Level
Science Context				
Provenance				
Project Organization and Sharing				
Reproducible Code				

Skills and Student Learning Outcomes

What to take away from this workshop

Category	Skill	Foundational	Second Level	Third Level
Science Context	1. Evaluating prior research 2. Acknowledgement 3. Intellectual Property			
Provenance	1. Creating Data 2. Using Data 3. Version Control 4. Documenting Versions			
Project Organization and Sharing	1. Sharing 2. Storage 3. Detailing 4. Balancing Reproducibility with Compliance			
Reproducible Code	1. Testing Code 2. Coding notebooks 3. Commenting 4. Documenting 5. Computational Environment			

Pedagogical Foundation

A systematic approach grounded in revised Bloom's Taxonomy

Revised Bloom's

Cognitive Process	Knowledge Types
Remember	Factual
Understand	Conceptual
Apply	Procedural
Analyze	Metacognitive
Evaluate	
Create	

Pedagogical Foundation

A systematic approach grounded in revised Bloom's Taxonomy

Revised Bloom's

Cognitive Process	Knowledge Types	Example	
Remember	Factual	**Category:**	Provenance
Understand	Conceptual	**Skill:**	Creating Data
Apply	Procedural	**Outcome:**	Understand spatial data structures
Analyze	Metacognitive	**Blooms:**	Understand, Conceptual
Evaluate			
Create			

Detailing

Specific learning outcomes associated with detailing project-level information

Category	Skill	Foundational	Second Level	Third Level
Science Context	1. **Evaluating research** 2. Acknowledgement 3. Intellectual Property	(F1) identify the research question and the type of research study (e.g., observational/experimental) (F2) identify claims/conclusions of a study (F3) deconstruct research workflow	(N1) evaluate internal validity of a study (N2) critique a research workflow (N3) plan to reproduce / replicate research	(T1) design solutions to address validity concerns (T2) evaluate context of study in literature (T3) attempt to replicate research in new context (e.g., location, population)
Bloom's Taxonomy		(F1) Analyze, Conceptual (F2) Analyze, Conceptual (F3) Evaluate, Procedural	(S1) Evaluate, Conceptual (S2) Evaluate, Procedural (S3) Create, Metacognitive	(T1) Create, Metacognitive (T2) Evaluate, Conceptuals (T3) Apply, Procedural

Version Control

Specific learning outcomes associated with the version control skill

Category	Skill	Foundational	Second Level	Third Level
Provenance	1. Creating Data 2. Using Data **3. Version Control** 4. Documenting Versions	(F1) understand what version control is (F2) create a local record of changes for your data and/or code	(S1) identify if application has version control (S2) use version control software tool (e.g., GitHub) individually to track workflows (e.g., push-pull or commit-add)	(T1) create a version control process (T2) collaboratively use version control (e.g., use development branches or workflows that require pull requests)
Bloom's Taxonomy		(F1) Understand, Factual (F2) Apply, Procedural	(S1) Analyze, Conceptual (S2) Apply, Procedural	(T1) Create, Procedural (T2) Apply, Metacognitive

Detailing

Specific learning outcomes associated with detailing project-level information

Category	Skill	Foundational	Second Level	Third Level
Project Organization and Sharing	1. Sharing 2. Storage **3. Detailing** 4. Balance Reproducibility with Compliance	(F1) sharing documented details about data and/or code, methods, or variables (metadata) (F2) organized project directory structure	(N1) write a README file with project-level metadata and geographic extent (N2) index project contents	(T1) Document data with ISO-standard geospatial metadata (T2) Archive project with Dublin core metadata and detailed geographic extent
Bloom's Taxonomy		(F1) Apply, Procedural (F2) Apply, Procedural	(S1) Create, Factual (S2) Analyze, Procedural	(T1) Create, Factual (T2) Create, Factual

Teaching Reproducibility and Replicability

Doing Reproduction and Replications with Students

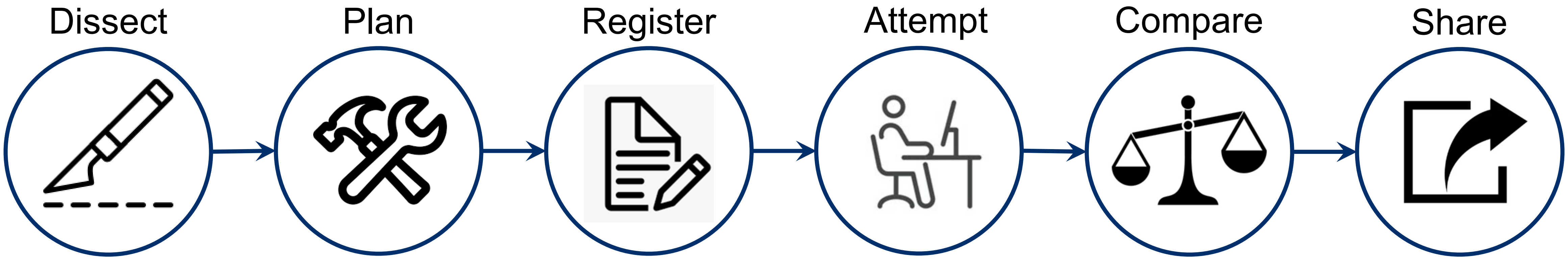
Summary Introduction

How we reproduce studies with students: settings

- Advanced undergraduate Open GIScience
- Graduate spatial statistics / GIS
- Independent research / thesis
- Summer research assistants
- Special reading/study groups

Summary Introduction

How we reproduce studies with students: process



Summary Introduction

How we reproduce studies with students: toolbox

- Complete Reproduction Studies
- Template Research Compendium
- Template Analysis Plan
- Course websites
- Suggested studies
- Project-based learning

Project Based Learning

Our pedagogical foundation

Project Based Learning Component	**Interpretation**	**Example Linkage to Reproductions and Replications**
1. Challenging Problem	*Meaningful question at appropriate level*	Critically evaluating the design and execution of prior studies
2. Sustained Inquiry	*Extended process of questioning*	Iterative evaluation and revision of work
3. Authenticity	*Real world context*	Engagement with peer-reviewed empirical research
4. Student Voice & Choice	*Decisions & implementation done by students*	Students contribute to project selection and lead project design/revision
5. Critique & Revision	*Give, receive & apply feedback*	Student to student consensus for replication design
6. Reflection	*Active reflection on learning and effectiveness of decisions*	Collective assessment of design decisions, analyses & unexpected challenges
7. Public Product	*Work shared beyond the classroom*	Share compendium of project design, data, code, results etc. for key decisions

Classroom Implementation I

Deconstruct a published study

Chakraborty, J. 2021.
Social inequities in the distribution of COVID-19: An intra-categorical analysis of people with disabilities in the U.S.
Disability and Health Journal 14:1-5.
https://doi.org/10.1016/j.dhjo.2020.101007

Data Sources

Methodology and Data Processing

Intermediary and Final Results

J. Chakraborty — *Disability and Health Journal 14 (2021) 101007*

Methods

Data on COVID-19 incidence were retrieved from the Johns Hopkins University Center for Systems Science and Engineering database[13] on August 1, 2020, for all counties in the continental U.S. This repository provides the most comprehensive and latest county-level COVID-19 data reported by the Centers for Disease Control and Prevention and state health departments, updated daily. The total number of COVID-19 cases in the 3108 counties of the continental U.S. (which excludes Alaska, Hawaii, and Puerto Rico) was 4,483,338 on the date this information was downloaded. The COVID-19 incidence rate, estimated as the number of confirmed cases per 100,000 people in each county, was used as the dependent variable for this study. The spatial distribution of this variable is depicted in Fig. 1, where counties in the continental U.S. are classified into five quintiles based on the COVID-19 incidence rate. Summary statistics for this dependent variable are included in the first row of Table 1.

Data on disability characteristics were obtained from the 2018 American Community Survey (ACS) five-year estimates. The ACS defines PwDs as members of the civilian non-institutionalized population who reported having serious self-care, hearing, vision, independent living, ambulatory, and/or cognitive difficulties on the ACS form. The ACS disability estimates allow disaggregation of PwDs based on five socio-demographic categories (race, ethnicity, poverty status, age, and biological sex) that were used for this intra-categorical analysis. County percentages for each disability subgroup were calculated by dividing the number of PwDs in each subgroup by the total civilian non-institutionalized population relevant to the variable category. The names and descriptive statistics for these explanatory variables are provided in Table 1.

Bivariate Pearson product-moment correlations were first used to measure statistical associations between COVID-19 incidence rate and each disability variable. Generalized estimating equations (GEEs) were then used for a multivariate analysis of disability subgroups within each socio-demographic category. GEEs extend the generalized linear model to accommodate clustered data,[14] in addition to relaxing several assumptions of traditional regression (i.e., normality).

For estimating a GEE, clusters of observations must be defined based on the assumption that observations within a cluster are correlated, while observations from different clusters are independent.[15] A combination of two different approaches were utilized to define county clusters for this study. The state in which a county is located was first used to account for potential correlation in counties within the same state, because of similar COVID-19 response and testing policies, socio-cultural systems, and health-care system characteristics[16–18] that imply similarities in counties within a given state and differences between states. Since the use of states as the only clustering variable potentially ignores intra-state and regional geographic variations in COVID-19 outcomes, a second approach based on identifying significant clusters of COVID-19 cases was incorporated. Specifically, SatScan[19] software was used to implement a spatial scan statistic based on the Poisson model, determine spatial clusters, and estimate relative risk (RR) for COVID-19 incidence rates at the county level. A similar methodology was recently employed by Desjardins et al.[20] to detect space-time clusters of COVID-19 cases in the continental U.S. The RR is defined as the estimated risk at a given location divided by the risk outside of the location (or, everywhere else). If a county has a RR of 3.0, for example, then the population within that county are three times more likely to be exposed to COVID-19. All U.S. counties were classified into six groups based on the estimated RR values (<1.0, 1.00–1.99, 2.00–2.99, 3.00–3.99, 4.00–4.99, and 5.0 or more). The use of both states ($n = 49$) and RR groups ($n = 6$) for the GEE cluster definition resulted in a total of 102 clusters, with the number of counties per cluster ranging from 1 to 245.

GEEs also require the specification of an intra-cluster dependency correlation matrix.[15] The 'exchangeable' correlation matrix was selected for the results reported here, since this specification yielded the best statistical fit based on the QIC (quasi-likelihood under the independence) model criterion. For each GEE, the normal, gamma, and inverse Gaussian distributions with logarithmic and identity link functions were explored. The gamma distribution with logarithmic link function was chosen for all GEEs since this model specification provided the lowest QIC value.

Since PwDs were disaggregated separately based on five socio-demographic characteristics (i.e., race, ethnicity, poverty status, age, and biological sex), five different GEE models were utilized. Each GEE included all disability subgroups relevant to that socio-demographic category. Finally, potential multicollinearity among the variables was also examined based on variance inflation factor,

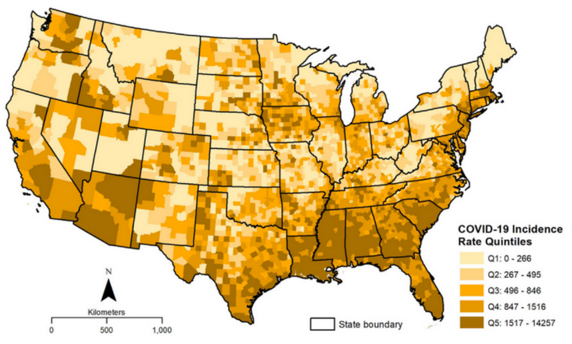
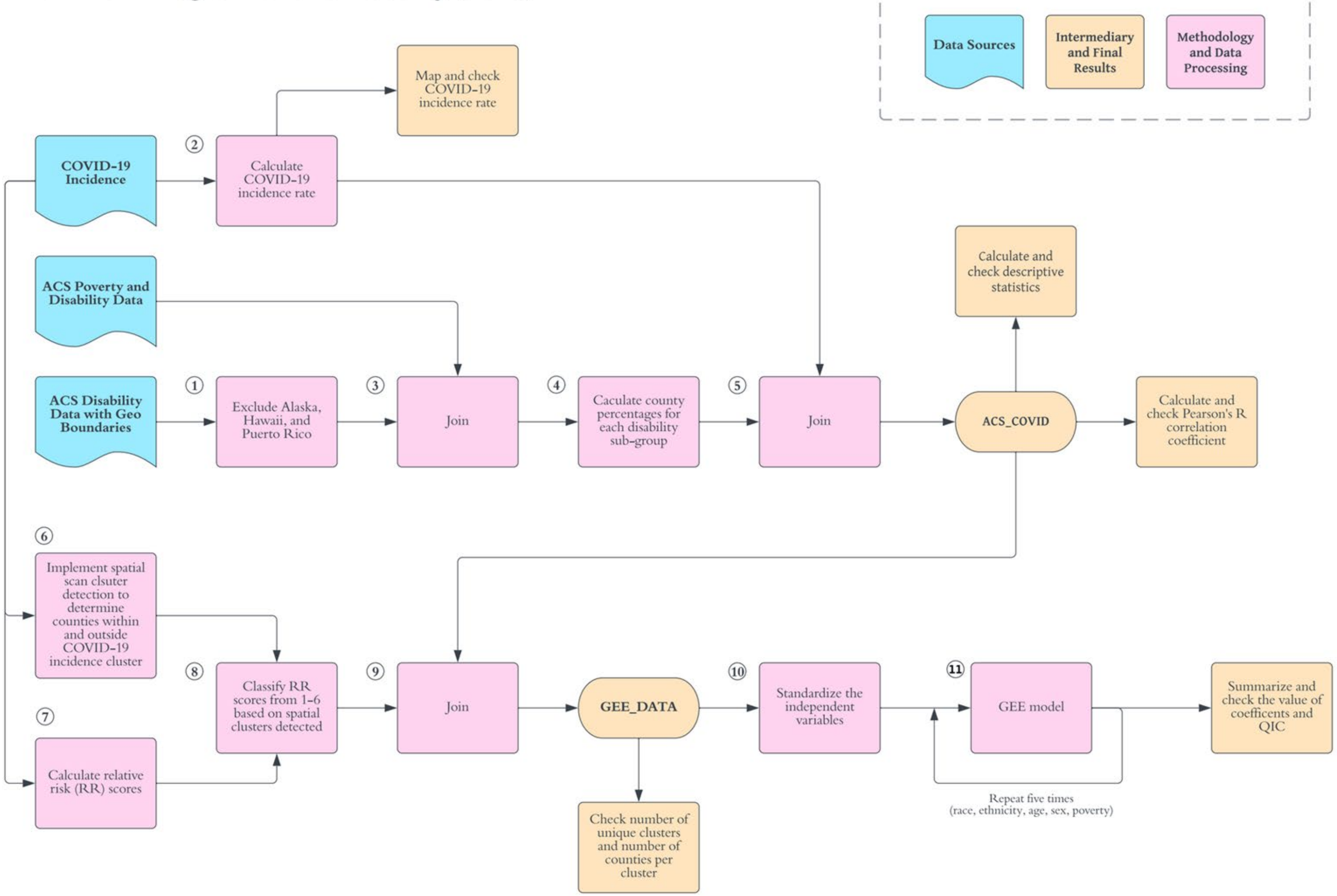


Fig. 1. County level distribution of COVID-19 incidence rate (cases per 100,000 people) in the continental USA, August 1, 2020.

Classroom Implementation I

Deconstruct a published study

Workflow Diagram of Chakraborty (2020)



Data Sources

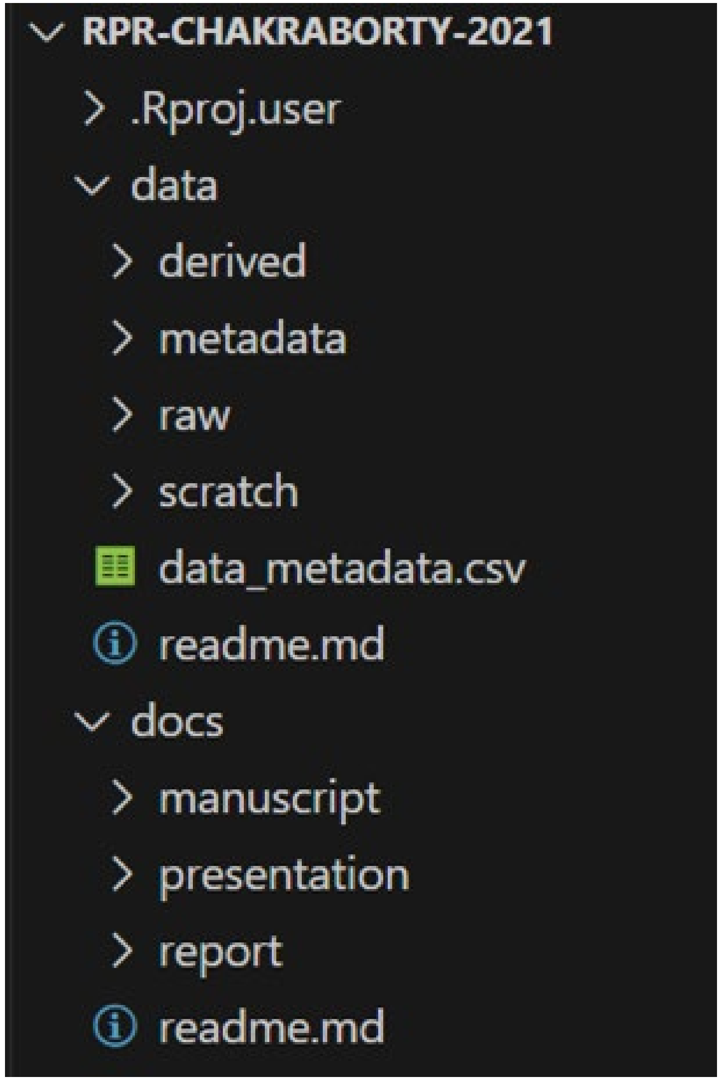
Methodology and Data Processing

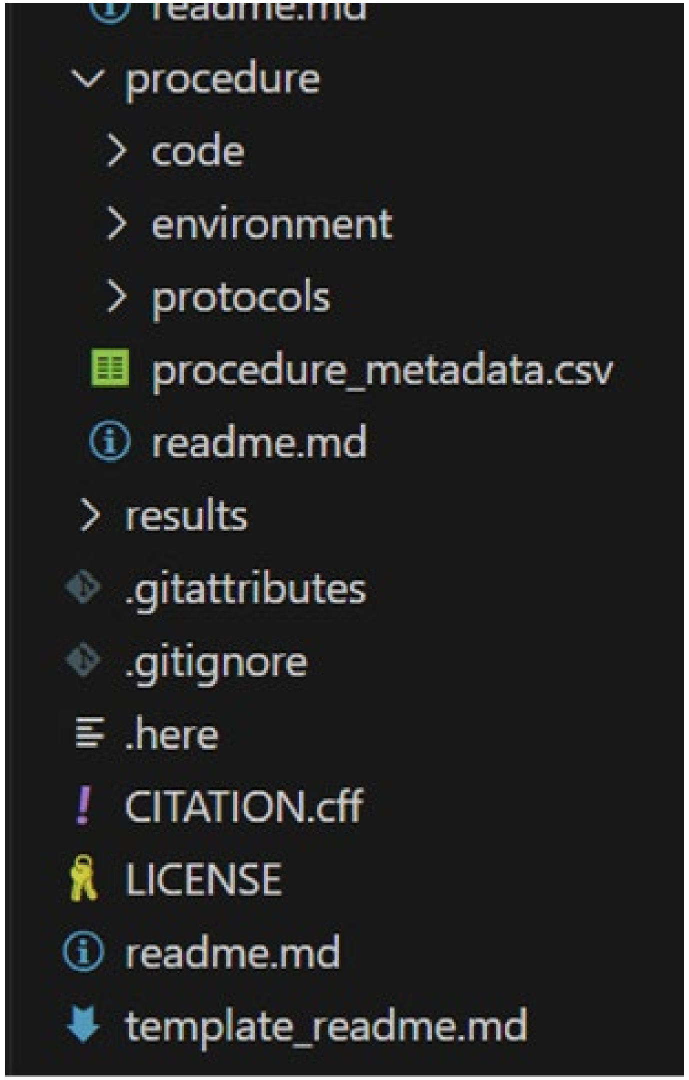
Intermediary and Final Results

Classroom Implementation II

Plan a reproduction attempt with a research compendium template

1. Directory structure
2. Version Control
3. Metadata:
   - Study
   - Data Sources



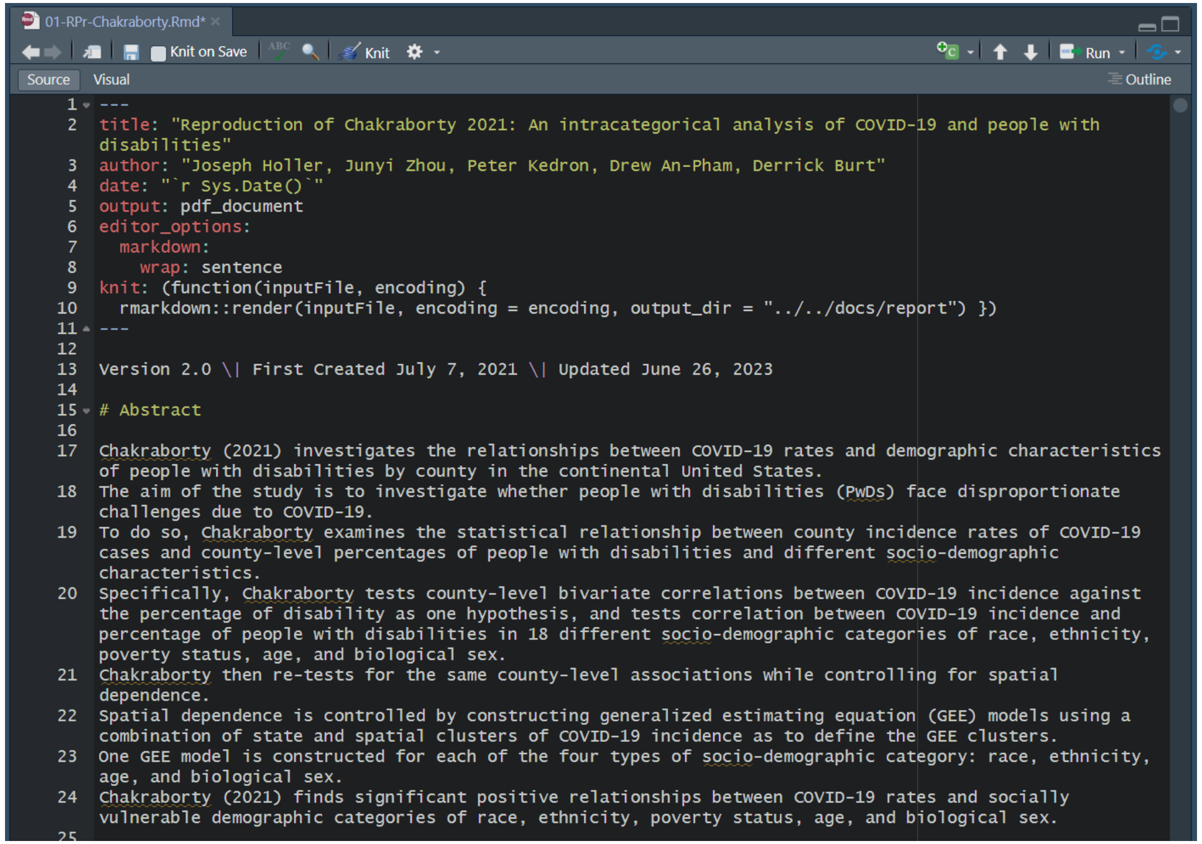


Classroom Implementation II

Plan a reproduction attempt with a research compendium template

1. Analysis Plan in Computational Notebook

➡ PDF Preregistration


01-RPr-Chakraborty.Rmd*

Knit on Save | Knit | Run

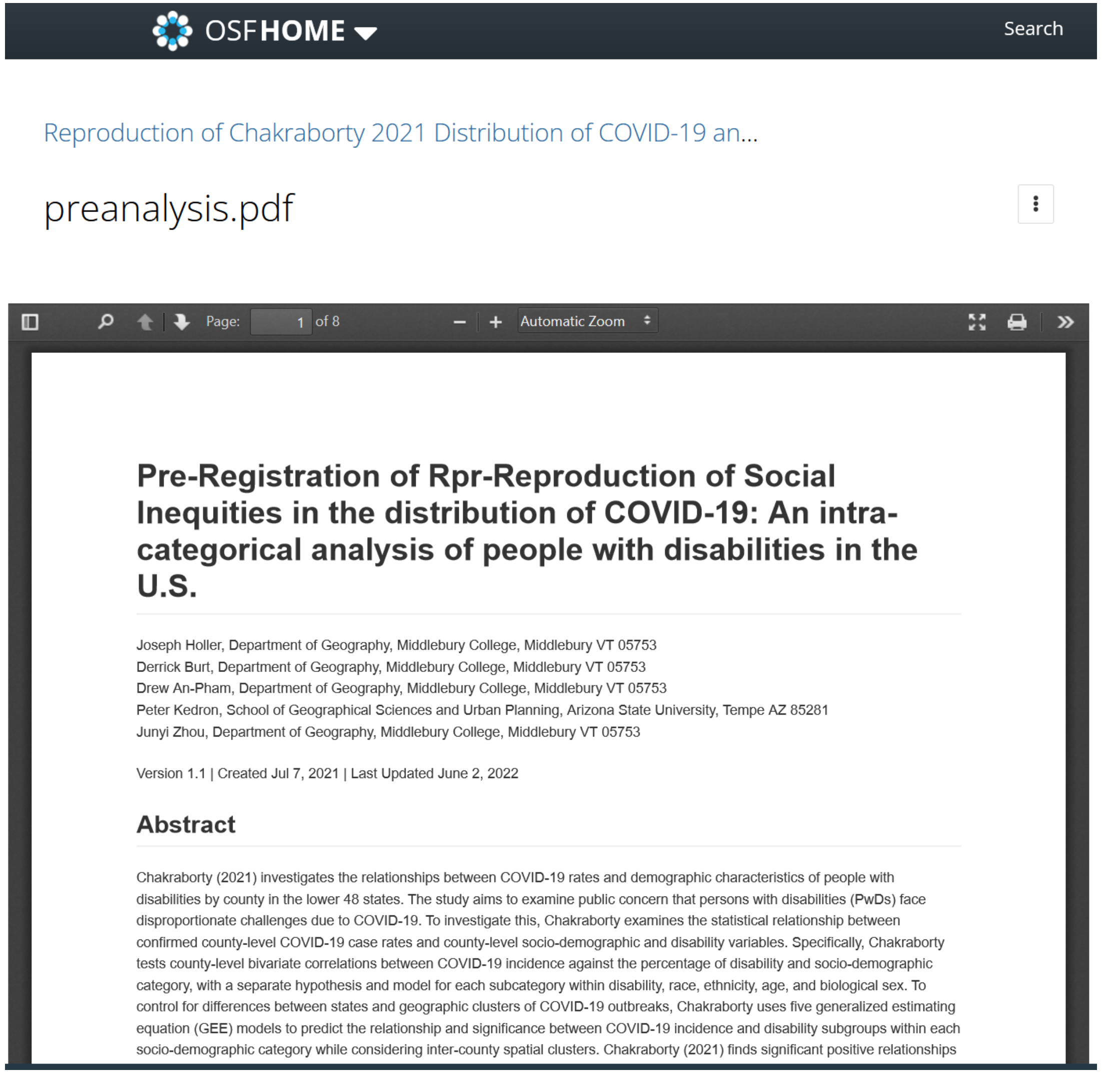
Source | Visual | Outline

```
---
title: "Reproduction of Chakraborty 2021: An intracategorical analysis of COVID-19 and people with disabilities"
author: "Joseph Holler, Junyi Zhou, Peter Kedron, Drew An-Pham, Derrick Burt"
date: "`r Sys.Date()`"
output: pdf_document
editor_options:
  markdown:
    wrap: sentence
knit: (function(inputFile, encoding) {
  rmarkdown::render(inputFile, encoding = encoding, output_dir = "../../docs/report") })
---

Version 2.0 \| First Created July 7, 2021 \| Updated June 26, 2023

# Abstract

Chakraborty (2021) investigates the relationships between COVID-19 rates and demographic characteristics of people with disabilities by county in the continental United States.
The aim of the study is to investigate whether people with disabilities (PwDs) face disproportionate challenges due to COVID-19.
To do so, Chakraborty examines the statistical relationship between county incidence rates of COVID-19 cases and county-level percentages of people with disabilities and different socio-demographic characteristics.
Specifically, Chakraborty tests county-level bivariate correlations between COVID-19 incidence against the percentage of disability as one hypothesis, and tests correlation between COVID-19 incidence and percentage of people with disabilities in 18 different socio-demographic categories of race, ethnicity, poverty status, age, and biological sex.
Chakraborty then re-tests for the same county-level associations while controlling for spatial dependence.
Spatial dependence is controlled by constructing generalized estimating equation (GEE) models using a combination of state and spatial clusters of COVID-19 incidence as to define the GEE clusters.
One GEE model is constructed for each of the four types of socio-demographic category: race, ethnicity, age, and biological sex.
Chakraborty (2021) finds significant positive relationships between COVID-19 rates and socially vulnerable demographic categories of race, ethnicity, poverty status, age, and biological sex.
```


OSFHOME | Search

Reproduction of Chakraborty 2021 Distribution of COVID-19 an...

preanalysis.pdf

Page: 1 of 8 | Automatic Zoom

Pre-Registration of Rpr-Reproduction of Social Inequities in the distribution of COVID-19: An intra-categorical analysis of people with disabilities in the U.S.

Joseph Holler, Department of Geography, Middlebury College, Middlebury VT 05753
Derrick Burt, Department of Geography, Middlebury College, Middlebury VT 05753
Drew An-Pham, Department of Geography, Middlebury College, Middlebury VT 05753
Peter Kedron, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe AZ 85281
Junyi Zhou, Department of Geography, Middlebury College, Middlebury VT 05753

Version 1.1 | Created Jul 7, 2021 | Last Updated June 2, 2022

Abstract

Chakraborty (2021) investigates the relationships between COVID-19 rates and demographic characteristics of people with disabilities by county in the lower 48 states. The study aims to examine public concern that persons with disabilities (PwDs) face disproportionate challenges due to COVID-19. To investigate this, Chakraborty examines the statistical relationship between confirmed county-level COVID-19 case rates and county-level socio-demographic and disability variables. Specifically, Chakraborty tests county-level bivariate correlations between COVID-19 incidence against the percentage of disability and socio-demographic category, with a separate hypothesis and model for each subcategory within disability, race, ethnicity, age, and biological sex. To control for differences between states and geographic clusters of COVID-19 outbreaks, Chakraborty uses five generalized estimating equation (GEE) models to predict the relationship and significance between COVID-19 incidence and disability subgroups within each socio-demographic category while considering inter-county spatial clusters. Chakraborty (2021) finds significant positive relationships

Classroom Implementation III

Attempt a reproduction, addressing unexpected issues, and assessing outcomes

1. Analysis Plan in Computational Notebook
   - ⇨ PDF Preregistration
2. Attempt Reproduction
   - + code blocks
   - + check results
   - + unplanned deviations

```
American Community Survey (ACS) data for sociodemographic subcategories of people with
disabilities can be accessed by using the `tidycensus` package to query the Census API. This
requires an API key which can be acquired at
[api.census.gov/data/key_signup.html](https://api.census.gov/data/key_signup.html).

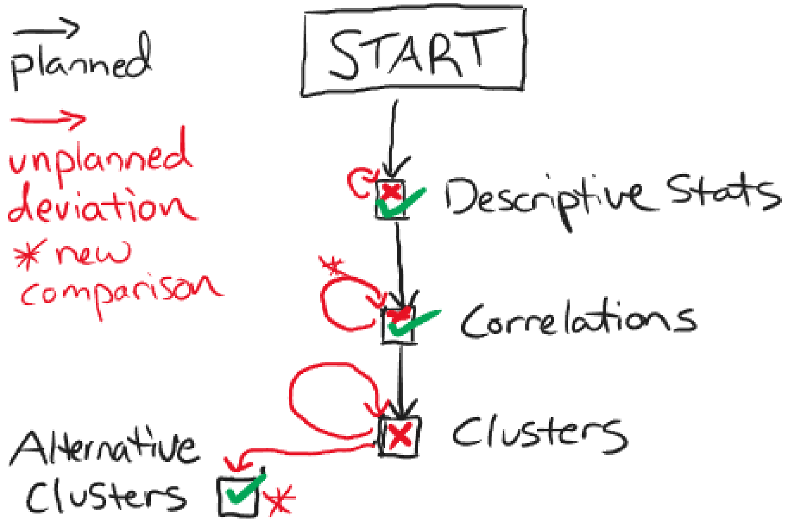
```{r API-Load-ACS, eval=FALSE}
# If you wish to use a census API key, run the census_api_key() function in the console

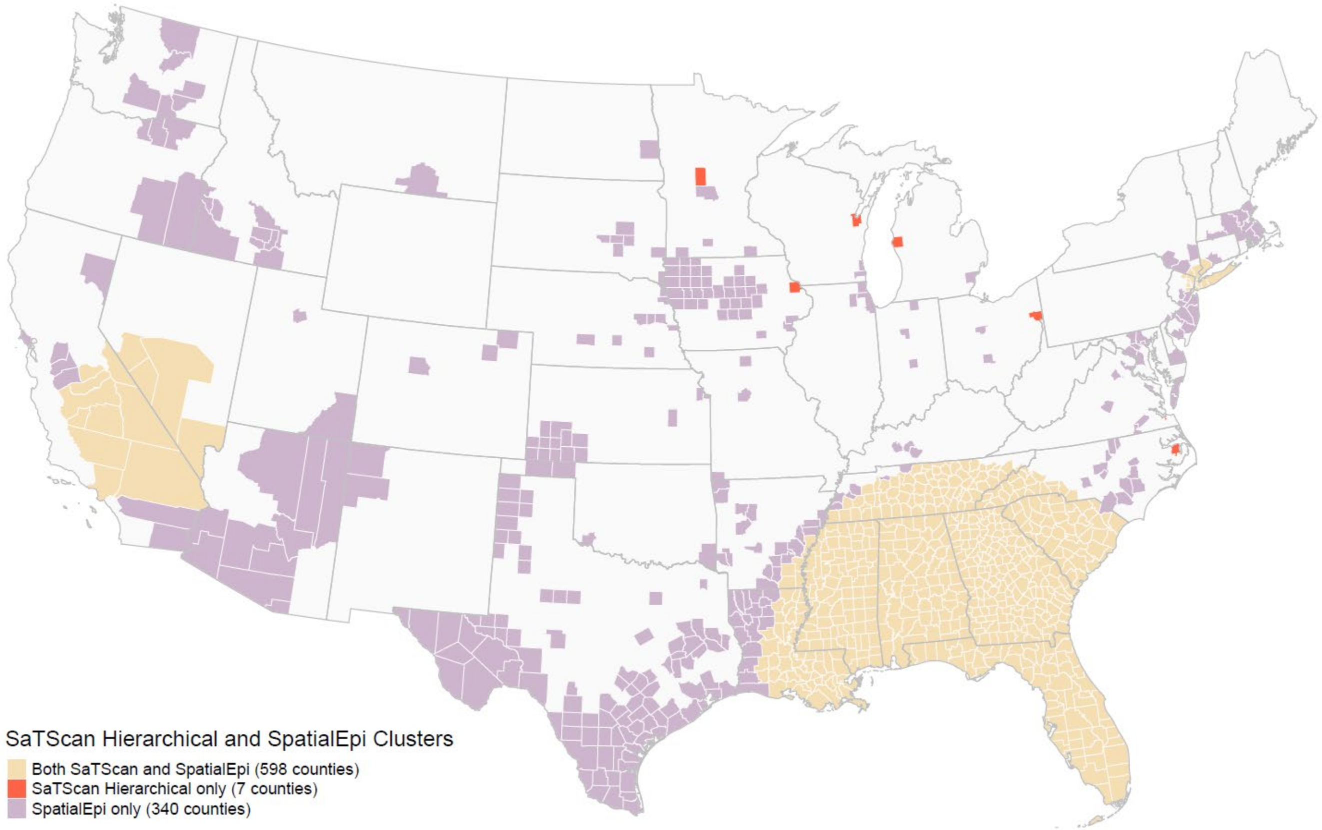
# Query disability demographic data with geographic boundaries
acs <- get_acs(
  geography = "county",
  table = "S1810",
  year = 2018,
  output = "wide",
  cache_table = TRUE,
  geometry = TRUE,
  keep_geo_vars = TRUE
)
```

adding tidycensus code to query ACS data to the data sources section
```

Classroom Implementation III

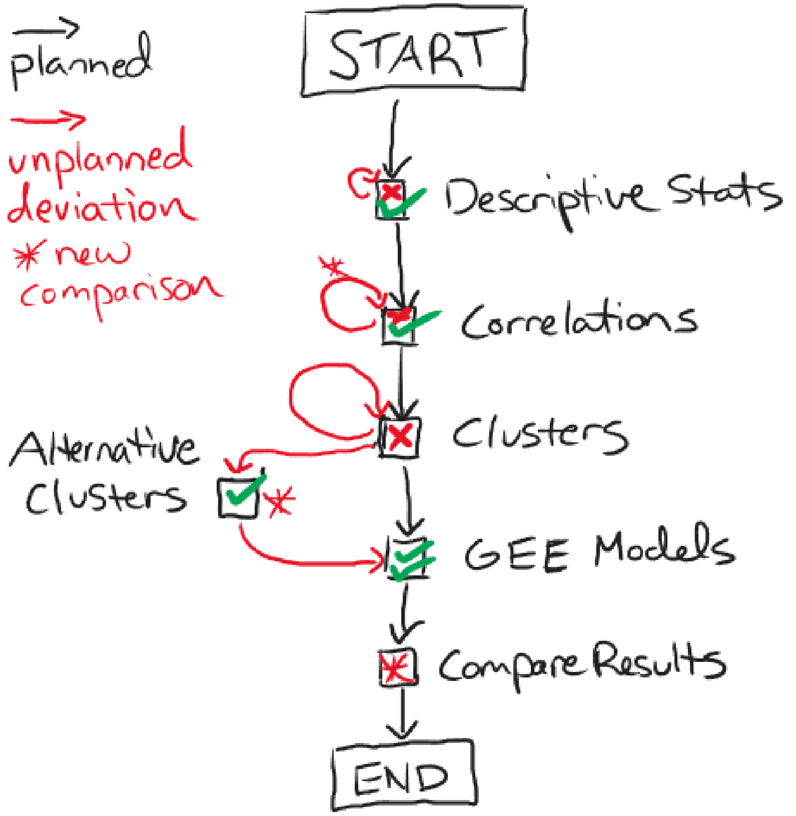
Attempting a reproduction, addressing unexpected issues, and assessing outcomes

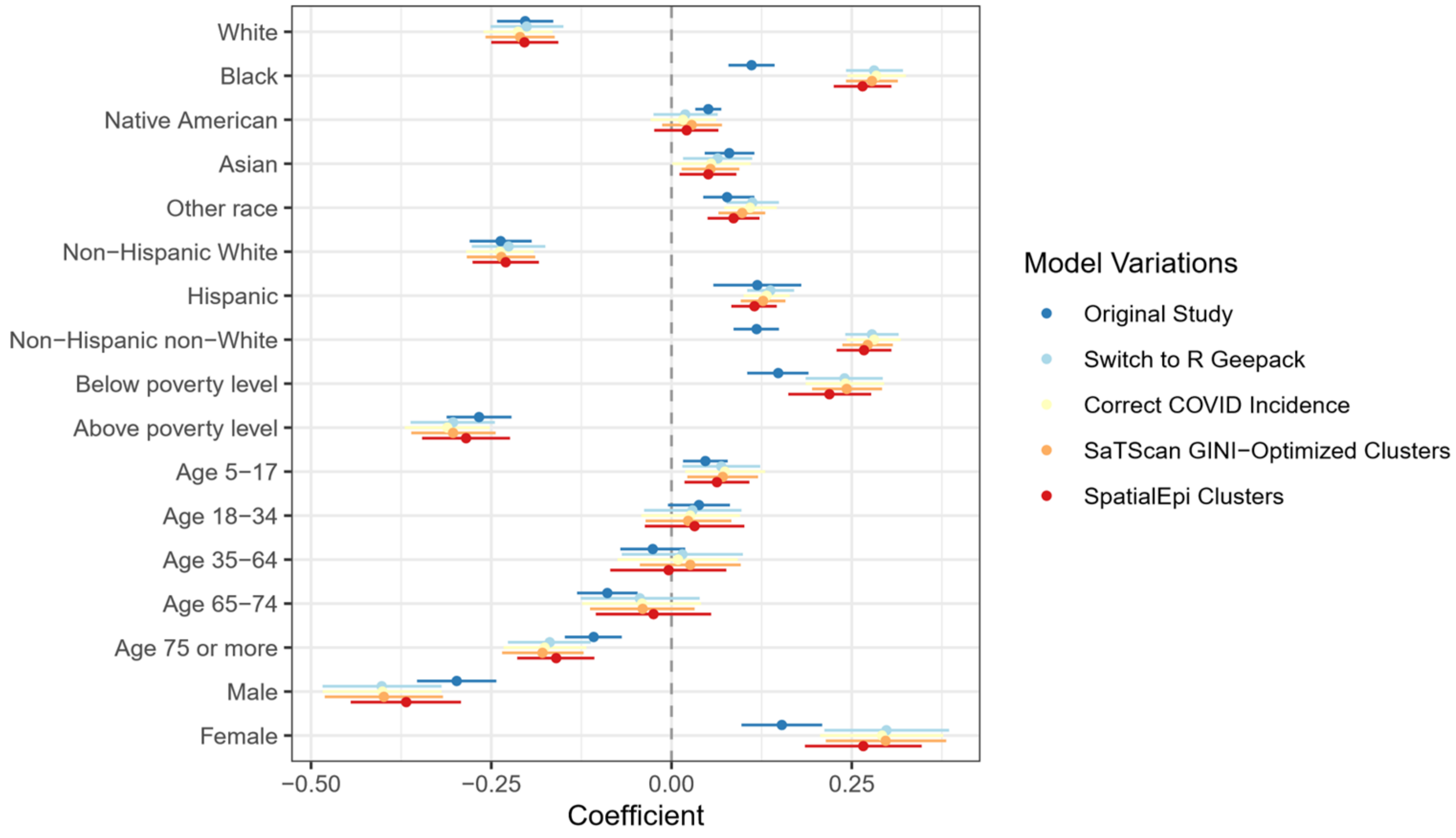




Classroom Implementation III

Attempting a reproduction, addressing unexpected issues, and assessing outcomes

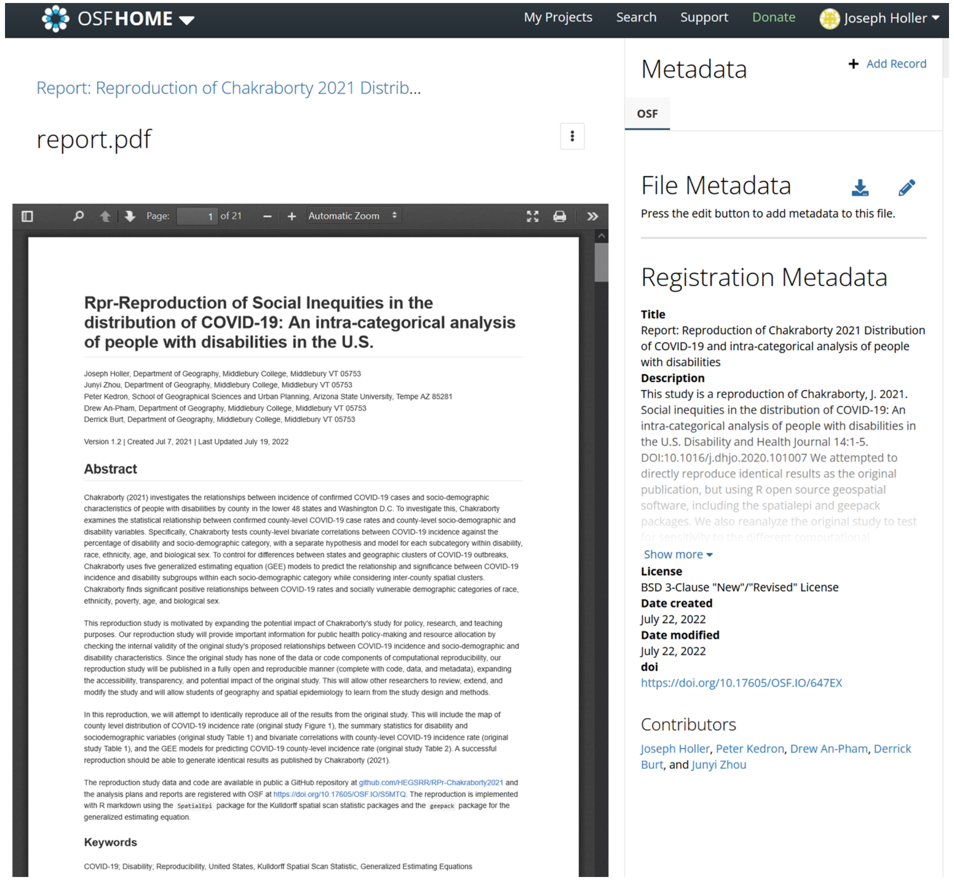




Classroom Implementation IV

Create and share a reproducible, public report

1. Analysis Plan in Computational Notebook
   - ⇨ PDF Preregistration
2. Attempt Reproduction
   - ✚ code blocks
   - ✚ check results
   - ✚ unplanned deviations
3. Analysis Report
   - ✚ output tables + figures
   - ✚ discuss unplanned deviations
   - ✚ conclusions
     - ⇨ PDF Registered Report
     - ⇨ HTML Website



Outcomes and Resources

Eleven studies and counting…

Original Study	Level	Pre-analysis	Compendium	Report	Publication
Malcomb et al 2014	Undergraduate	GitHub	GitHub; OSF	—	—-
Kang et al 2020	Undergraduate	GitHub	GitHub; OSF	OSF	*International Journal of Health Geographics* - In Review
Mollalo et al 2020	Graduate	GitHub	GitHub; OSF	GitHub	*Geographical Analysis*
Saffary et al 2020	Graduate	GitHub	GitHub; OSF	GitHub	*Geographical Analysis*
Vijayan et al 2021	Graduate	GitHub	GitHub; OSF	GitHub	*Geographical Analysis*
Chakrabory 2021	Undergraduate	OSF	GitHub; OSF	OSF	
DiMaggio et al 2020	Graduate	OSF	GitHub; OSF	OSF	*Annals of Epidemiology*
Speilman et al 2020	Undergraduate	OSF	GitHub; OSF	OSF	
Maldonado	Undergraduate	Github - Private	GitHub - Private	GitHub - Private	*Journal of Immigrant and Minority Health* - Submit Fall 2024
Kodros	Graduate	Github - Private	GitHub - Private	GitHub - Private	
Brodie	Graduate	Github - Private	GitHub - Private; OSF - Private	Github - Private	*Nature* - Submit Fall 2024

Outcomes and Resources

Workshop Website

hegsrr.github.io/Workshop-SDSS-2024/

- 5 Peer-reviewed Publications
- 11 Reproduction and Replication Studies
- 2 Surveys of Researcher Practices with interactive data visualizations
- **Research Compendium Template**
- Manual In Development
- Course Syllabi
- 9 RAs Mentored
- ~75 Students Engaged in R&R Studies

Key Ideas for Teaching Reproducibility and Replicability

What to take away from this workshop

(1) Teach more than data and code sharing.
Link R&R to the epistemology of science and open science

(2) Attempt reproductions and replications with your students
Pedagogically rich way to teach GIScience, spatial data science, and topical knowledge

(3) Reproduction attempts create intrinsic and extrinsic rewards
Reproduction attempts are rewarding for students, improve learning outcomes, produce publications

(4) Use (and improve) our open educational materials
Templates, past reproductions, teaching materials

Workshop Website

hegsrr.github.io/Workshop-SDSS-2024/

An Open Invitation

Please reach out to us.
We want to collaborate on R&R research.
We want to help those who are interested adopt this approach

Funding Support from NSF BCS-2049837