

Practicing and Teaching Reproducibility and Replicability in the Human-Environment and Geographical Sciences

Joseph Holler – Middlebury College

Peter Kedron – Arizona State University

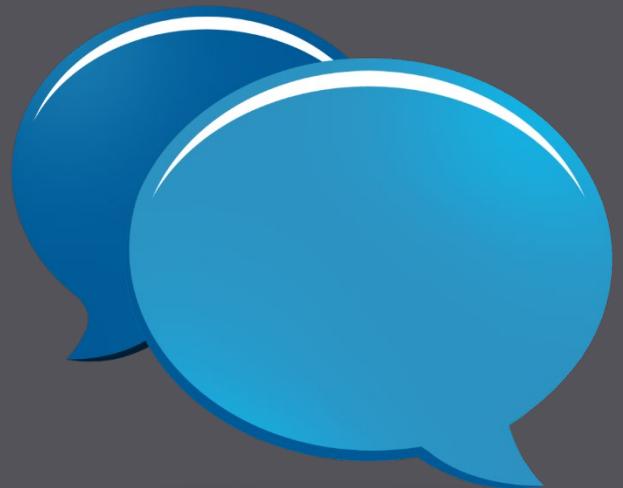
Emily Zhou – Middlebury College

Support

- NSF BCS-2049837
Transforming Theory-Building and STEM Education Through Reproductions and Replications in the Geographical Sciences
 - -UCGIS
- Geospatial Software Institute - COVID-19 Fellows
University of Illinois Urbana-Champaign
National Science Foundation OAC-1743184



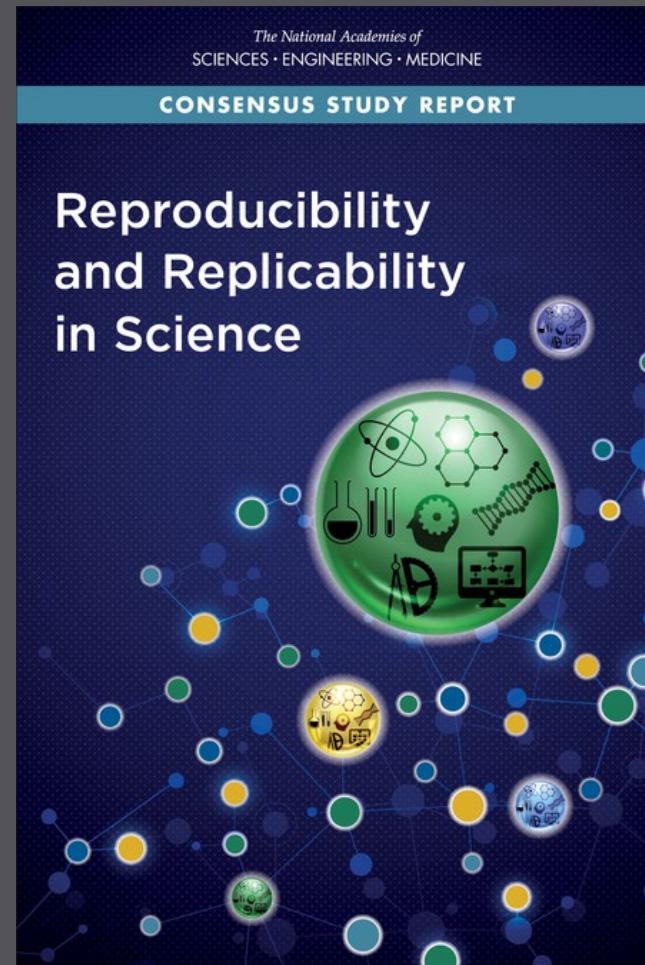
Introductions



- Name
- Institution
- Research or teaching project related to R&R

Reproducibility & Replicability (R&R)

- **Reproducibility:** Obtaining consistent results using the same input data, computational steps, methods and code, and conditions of analysis
- **Replicability:** Obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data



Practicing R&R

- Adopting reproducible research practices
- Conducting reproduction / replication studies

Teaching R&R

- Teaching competencies for practicing reproducible research
- Conducting reproduction and replication studies with students in...
 - Methods courses
 - Independent or small group studies
 - Research assistantships
- Teaching geographic concepts and methods through reproduction and replication studies

Today's Seminar Objectives

1. Overview of R&R in Human-Environment and Geographical Sciences
2. Research compendium in Git / GitHub
3. Plan and register a reproduction / replication study
4. Execute a computational notebook in Rstudio / R Markdown
 - Share & discuss experience & vision for R&R
 - We welcome feedback on all materials shared today!
Email, GitHub Issue, or Pull Request!

Our Research Objectives

1. Assess reproducible research practices in the geographical sciences and identify barriers to reproducibility.
2. Assess the credibility and generalizability of recent high-impact HEGS research by conducting reproductions and replications.
3. Establish and evaluate a pedagogy that uses reproductions and replications to improve student learning and STEM competencies.

State of the Research: Year One

- Curriculum: <https://gis4dev.github.io> *find learning resources here!
- Seven reproduction / replication studies:
 - <https://osf.io/c5a2r/> and <https://github.com/HEGSRR>
- Kedron, P., and J. Holler. 2022. Replication and the search for the laws in the geographic sciences. *Annals of GIS* 28 (1):45–56.
- A Replication of Dimaggio Et Al. (2020) in Phoenix, AZ.
 - *Annals of Epidemiology*
- Moving Beyond Computation: Reproducing Geographical Analyses of COVID-19 to Assess and Improve the Validity of Research
 - *Geographical Analysis*

Findings from Reproductions

1. Every study can be improved: reproducibility can be constructive!
2. The basic components needed to reproduce studies were not available
3. Metadata was missing and seems to have not been closely examined by some authors
4. There is a clear need for true interdisciplinary work that balances domain knowledge with expertise in spatial analysis
5. Spatial analysis decisions were often weakly justified and not carefully examined
6. The selective inference problem makes it unclear how reliable even reproducible studies are

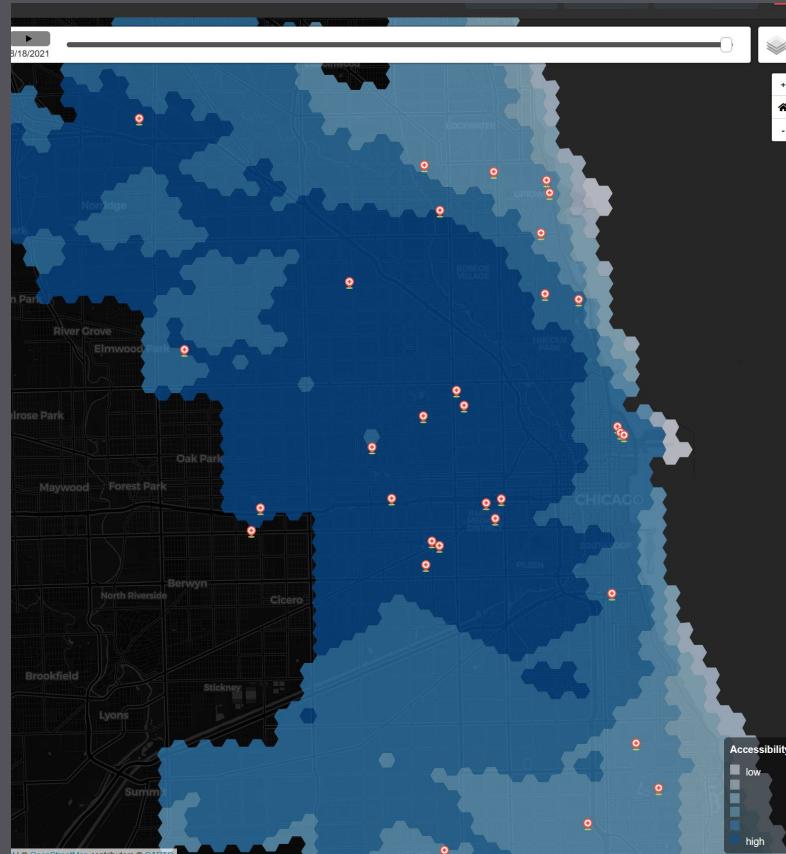
Part I: Overview of R&R in HEGS

Vignette: Spatial Accessibility in Illinois

R&R illustrated in context of COVID-19 pandemic

Vignette: Spatial Accessibility in Illinois

Kang, J. Y., A. Michels, F. Lyu, Shaohua Wang, N. Agbodo, V. L. Freeman, and Shaowen Wang. 2020. Rapidly measuring spatial accessibility of COVID-19 healthcare resources: a case study of Illinois, USA. *International Journal of Health Geographics* 19 (1):1–17.
doi.org/10.1186/s12942-020-00229-x.



Reproduce Studies with Jupyter Notebook on CyberGISX

- <https://cybergisxhub.cigi.illinois.edu/wherencovid-19/>



Open with CyberGISX

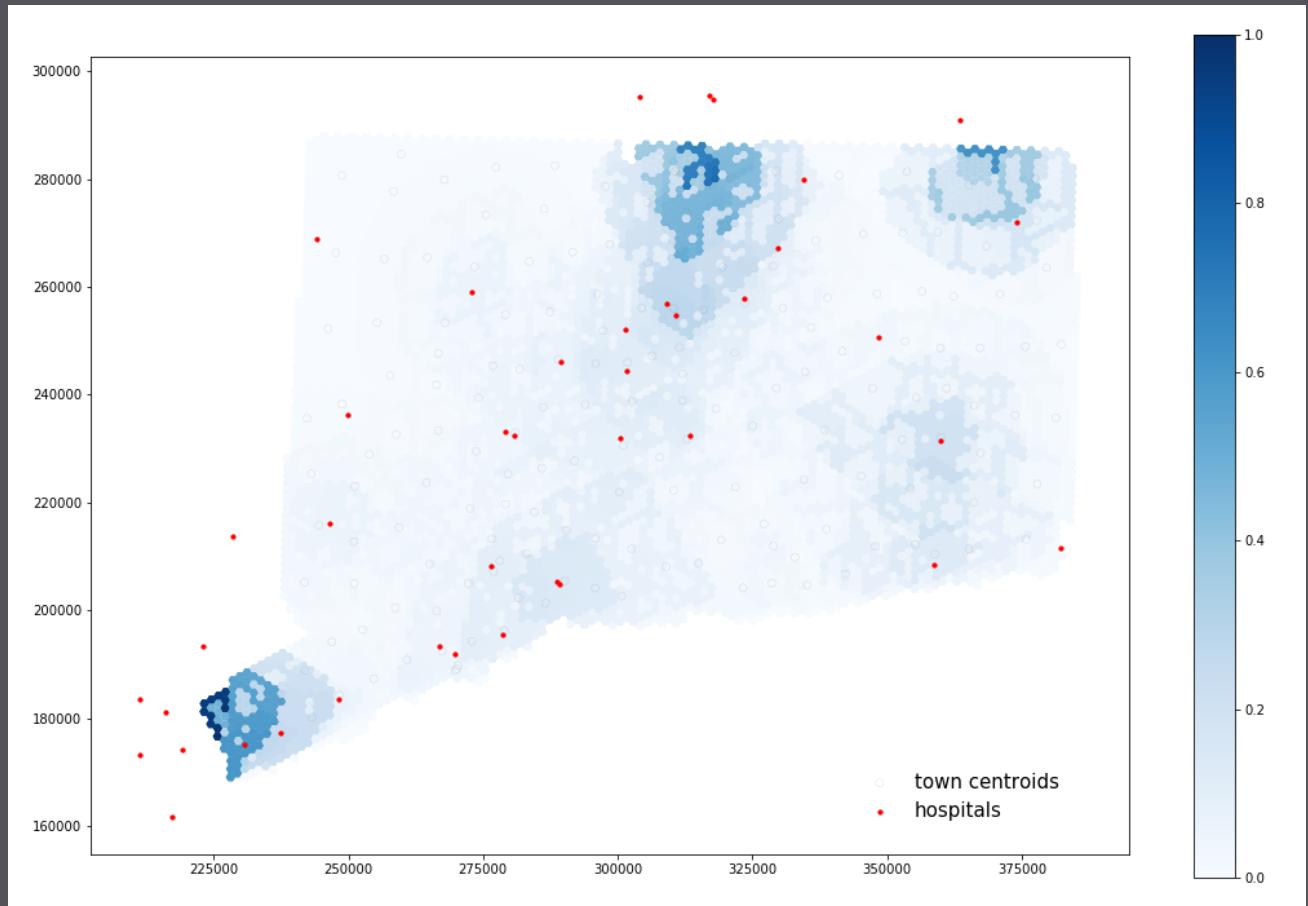
Generate and Plot Map of Hospitals

```
In [6]: m = folium.Map(location=[41.85, -87.65], tiles='cartodbdpositron', zoom_start=10)
for i in range(0, len(hospitals)):
    folium.CircleMarker(
        location=[hospitals.iloc[i]['Y'], hospitals.iloc[i]['X']],
        popup="{}{}\n{}{}\n{}{}".format('Hospital Name: ',hospitals.iloc[i]['Hospital'],
                                         'ICU Beds: ',hospitals.iloc[i]['Adult ICU'],
                                         'Ventilators: ', hospitals.iloc[i]['Total Vent']),
        radius=5,
        color='grey',
        fill=True,
        fill_opacity=0.6,
        legend_name = 'Hospitals'
    ).add_to(m)
legend_html = '''<div style="position: fixed; width: 20%; height: auto;
bottom: 10px; left: 10px;
solid grey; z-index:9999; font-size:14px;
">&ampnbsp Legend<br>'''
```

:[6]:

Fall 2020: Derrick Burt Reproduces and Replicates

- Reproduction of Chicago succeeds
- Reproduction of Illinois fails due to network size
- Replication in Connecticut with town-level COVID-19 data



January 2021: Kufre Udoh Reanalyzes

- Local python environment
- Update **osmnx** and **networkx** packages
- Replace **ego_graph()** with **subgraph()**
- See time comparisons!

```
In [6]: %%time
ego_ccc(G, hospitals['nearest_osm'][0], distances[2])
Wall time: 4.67 s
Out[6]: geometry
0 POLYGON ((-90.08384 38.60189, -90.08789 38.601...
<
In [7]: %%time
dijkstra_ccc(G, hospitals['nearest_osm'][0], distances[2])
Wall time: 880 ms
Out[7]: geometry
0 POLYGON ((-90.08384 38.60189, -90.08789 38.601...
<
In [8]: %%time
ego = []
for i in range(10):
    ego.append(ego_ccc(G, hospitals['nearest_osm'][i], distances[2]))
Wall time: 28.2 s
In [9]: %%time
dij = []
for i in range(10):
    dij.append( dijkstra_ccc(G, hospitals['nearest_osm'][i], distances[2]))
Wall time: 11 s
```

Spring 2021: Reproduce with class

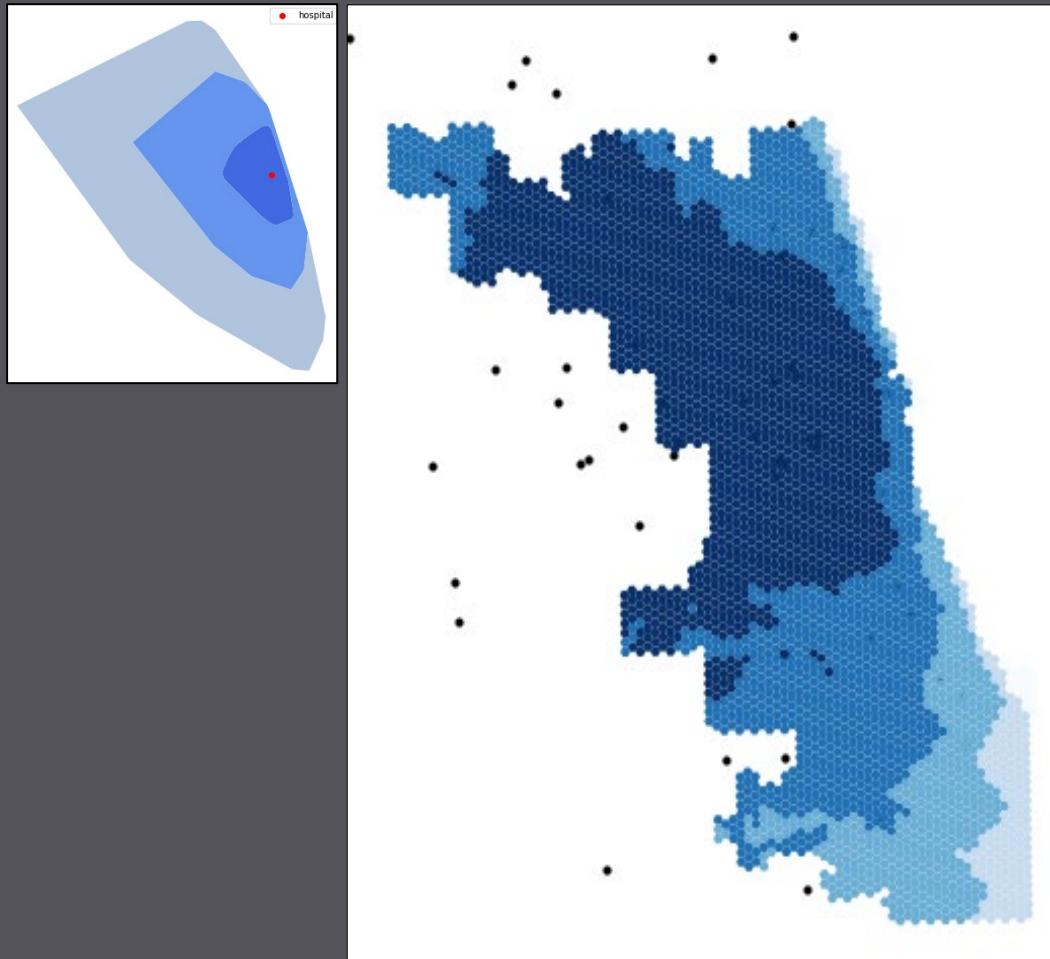
- Alexandermichels revises notebook to focus on Chicago
- Holler fixes bugs in notebook
- 17 Middlebury students reproduce the study for Chicago
- *Twenty* novice & experienced GIScientists have now run the notebook
- GitHub tracks networks & history of contributions

The screenshot shows a GitHub repository interface. On the left, a dark-themed sidebar displays a tree view of repository branches and pull requests. The main area shows a chronological list of commits from August 19, 2021, down to April 10, 2021. Commits are made by users like alexandermichels, josephholler, and derrickburt, often involving bug fixes and script updates. On the right, a list of contributors is shown, each with a small profile picture and their GitHub handle followed by their role: RP-Kang.

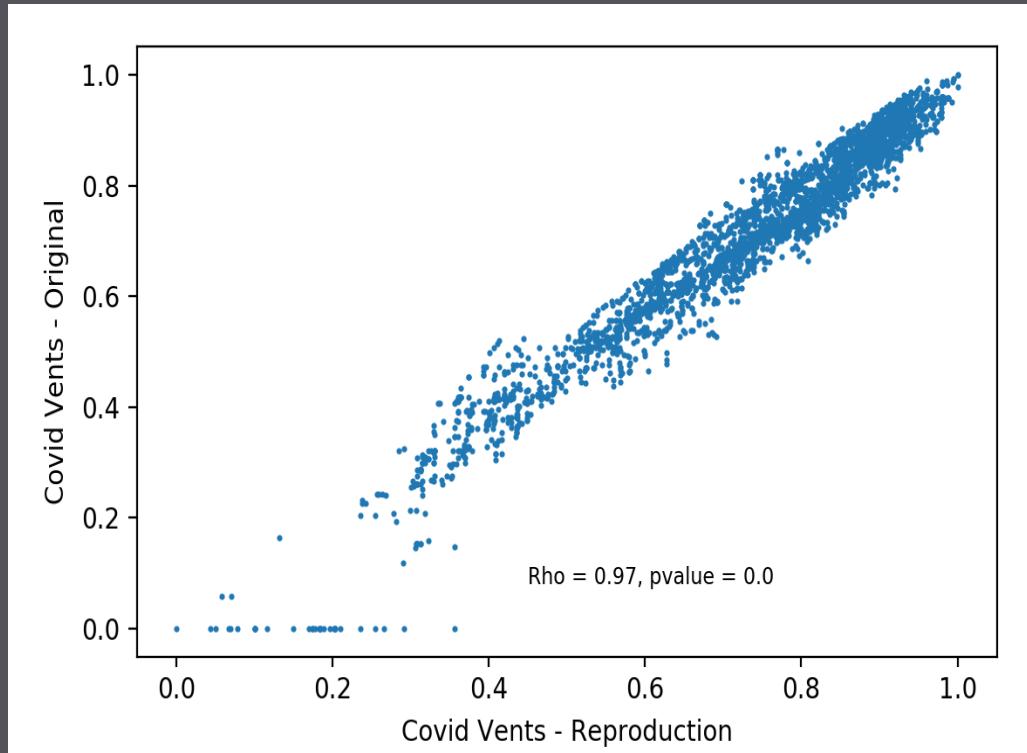
Contributor	Role
cybergis / COVID-19AccessibilityNotebook	
GIS4DEV / RP-Kang	
alandaux / RP-Kang	
avillanueva1005 / RP-Kang	
brookelaIRD / RP-Kang	
daptX / RP-Kang	
derrickburt / RP-Kang-Improvements	
HEGSRR / RPr-Kang-2020	
emmag725 / RP-Kang	
emmaclinton / RP-Kang	
evankilli / RP-Kang	
gsmarshall / RP-Kang	
hrigdon98 / RP-Kang	
jackson-mumper / RP-Kang	
jafreedman12 / RP-Kang	
kufreu / RP-Kang	
majacannavo / RP-Kang	
mtango99 / RP-Kang	
nicknonnen / RP-Kang	
sanjana-roy / RP-Kang	
stevenmontilla / RP-Kang	
vinfalardeau / RP-Kang	

Summer 2021: Derrick Burt Reanalyzes

- github.com/HEGSRR/RPr-Kang-2020
- Apply **HEGSRR Template & USGS Metadata Wizard 2.0**
- **Preprocess data** (-ventilators)
- Create **figures** in paper (histogram, classified map)
- **Credibility/Legibility:** Vignette/visualize a single hospital; check speed data (mostly NULL; new error)
- **Improve Accuracy:** 30mph urban speed; errors in new OSM data, edge effects
- **Improve Efficiency:** Replace three iterations of Dijkstra's Shortest Path for each hospital with one
- **Compare** original and reproduction results: correlation and scatterplot



Successful Computational Reproduction



- CyberGISX provides a **collaborative scholarly community** of GIScientists and students
- Future work:
 - New **osmnx** functions for **network pre-processing**
 - Improve construction of **catchment area polygons**
 - Use **geopandas spatial indices** for overlap analysis
 - Implement **social vulnerability analysis**
 - Create **data structure** to store hospital catchment areas as collection of hexagonal grids

Reproducibility & Replicability in the Geographical Sciences

Illustrated with COVID-19 Pandemic Research

Science

- Science builds explanatory structures, tells stories which are scrupulously tested to see if they are stories about real life.

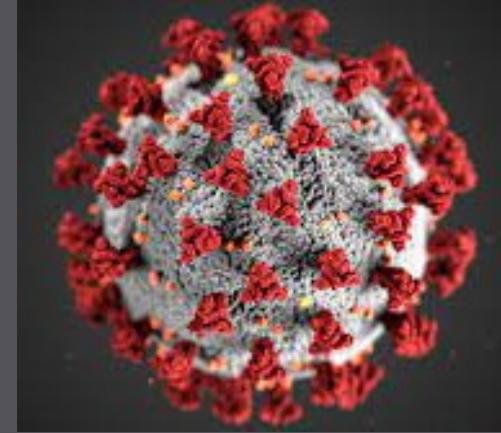
-Medawar 1967

- Science is about finding the most reliable way of thinking at the present level of knowledge.

-Rovelli 2014

Science Under Stress: COVID-19

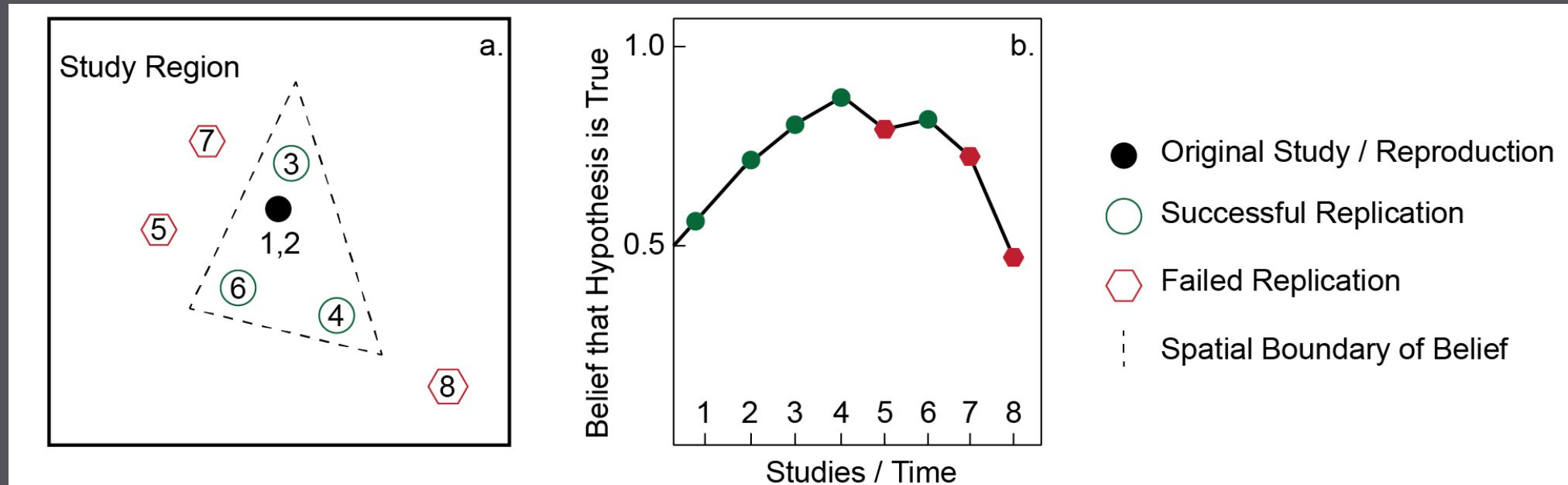
- Reproducibility Crisis
- Reproducibility rules @ Environmental Protection Agency & Office of Budget Administration
- Knowledge evolving rapidly
 - Open access preprint repositories medRxiv, bioRxiv
 - Accelerated peer review and publishing process
- Retractions from hasty science?
 - Hydroxychloroquine, Surgisphere, *The Lancet & New England Journal of Medicine*
 - *An alarming retraction rate for scientific publications on Coronavirus Disease*
(Yeo-Teh and Tang 2021)
- Crisis of public trust?



(How) Do we trust science?

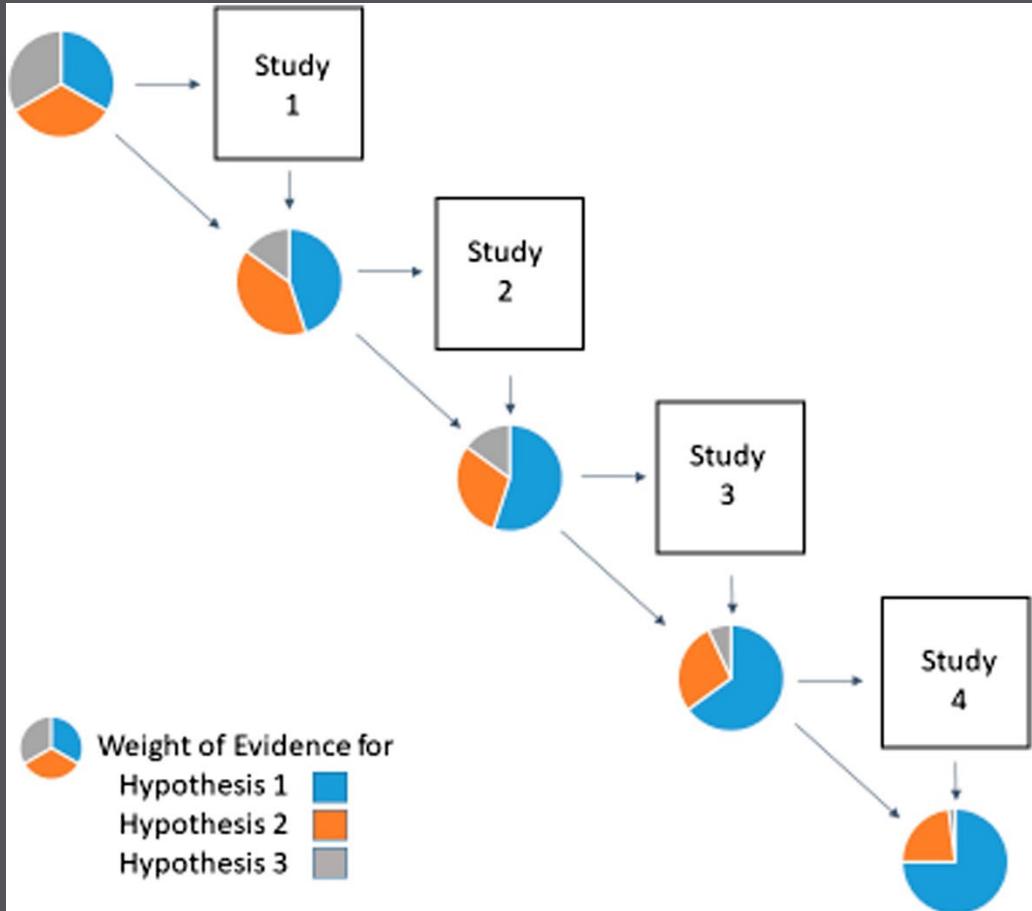
- Community with self-corrective mechanisms
 - Peer Review
 - Debate / test alternative hypotheses – Kuhn’s paradigm shifts
- Internal validity
- External validity
- Reliability
- Credibility

Spatial Autocorrelation, Heterogeneity & Scientific Discovery



- Kedron, P., and J. Holler. 2022. Replication and the search for the laws in the geographic sciences. *Annals of GIS* 28 (1):45–56.

Accumulation of Evidence



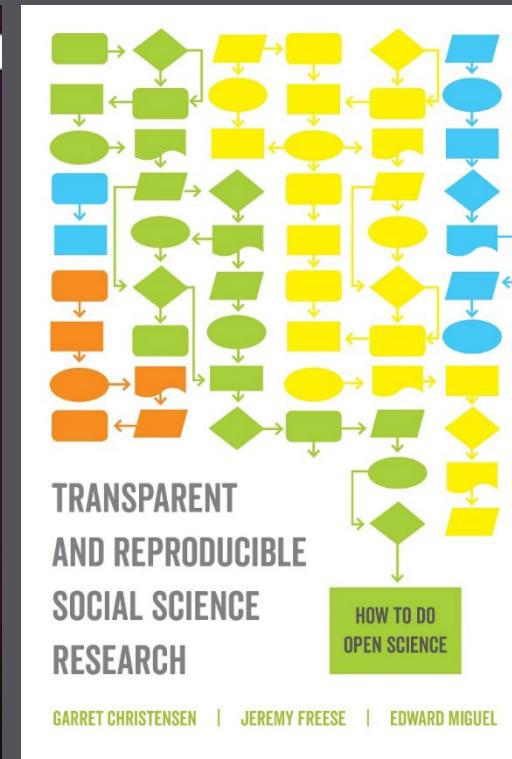
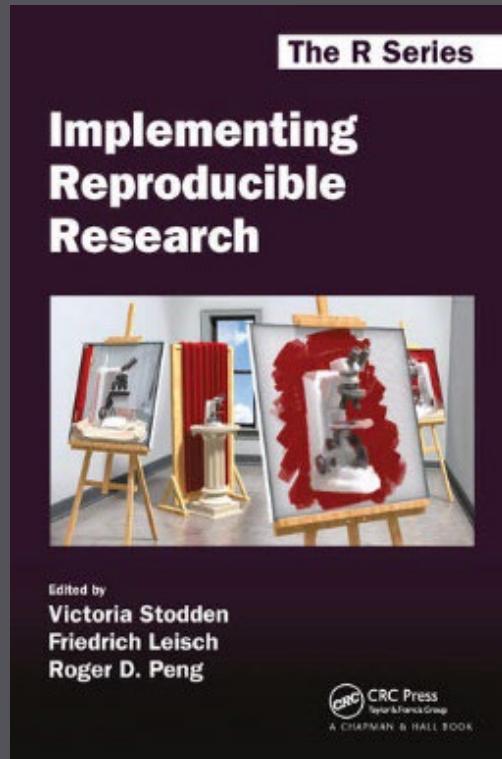
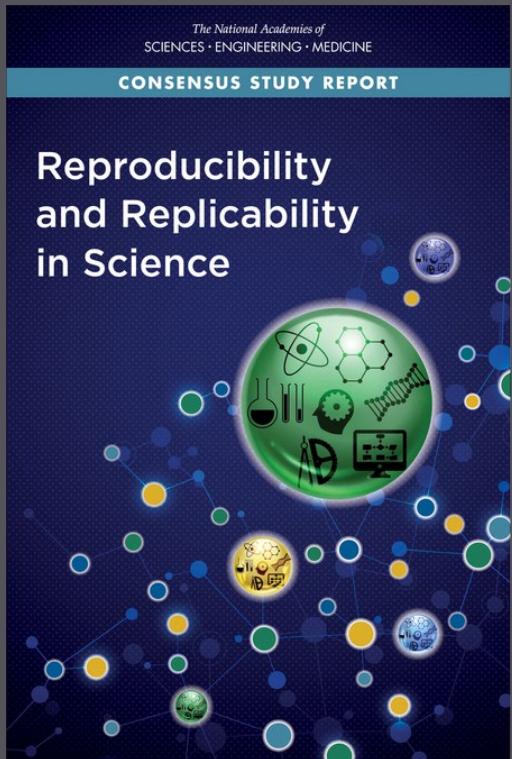
- Nichols, J. D., M. K. Oli, W. L. Kendall, and G. Scott Boomer. 2021. A better approach for dealing with reproducibility and replicability in science. *Proceedings of the National Academy of Sciences of the United States of America* 118 (7):1–5.

Reproduction for Credibility & Scientific Progress

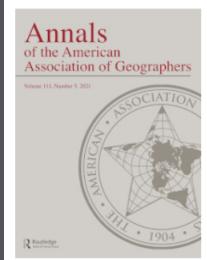
	Same Methods	Varied Methods
Same Data	Reproduction (Verification)	Reanalysis
Different Data	(Direct) Replication	Extension

Modified from Christensen, Freese and Miguel (2019, Table 9.1 pg 159)

Literature: Highlights



R&R in Geography: Highlights



**Annals
of the American
Association of Geographers**
Volume 111 Number 5, 2021

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/raag21>

**Introduction: Forum on Reproducibility and
Replicability in Geography**

Michael F. Goodchild, A. Stewart Fotheringham, Peter Kedron & Wenwen Li

To cite this article: Michael F. Goodchild, A. Stewart Fotheringham, Peter Kedron & Wenwen Li (2021) Introduction: Forum on Reproducibility and Replicability in Geography, *Annals of the American Association of Geographers*, 111:5, 1271-1274, DOI: [10.1080/24694452.2020.1806030](https://doi.org/10.1080/24694452.2020.1806030)

To link to this article: <https://doi.org/10.1080/24694452.2020.1806030>



o2r

Opening Reproducible Research is a
DFG-funded research project by
Institute for Geoinformatics (**ifgi**) and
University and Regional Library (**ULB**),
University of Münster, Germany

R&R in Geography: Highlights

- Computational reproducibility: AGILE, GIScience, VGI
 - Ostermann, Granell, o2r team
- Executable research compendia & containerizing processing environ.
 - o2r: Konkol, Kray, Nüst, Pebesma
- Open GIScience / Open Source GIS
 - S. Rey, D. Sui, A. Singleton, S. Spielman, C. Brunsdon, C. Farmer, H. Mitasova
- 5-star guide: open standards & metadata: Wilson et al
- Provenance / privacy / data quality: Tullis & Kar
- Epistemology and Reproducibility: Wainwright; D. Sui and P. Kedron
- Data Repository: Mei Po Kwan, D. Richardson

Practical & Computational Reproducibility

Many necessary conditions for Reproducibility



Improving R&R

Practices

- Preregister research plan
- Data
 - raw, preprocessed, and results
- Code
 - computational notebook
 - legibility
- Processing environment
 - Open source software
 - Specify or containerize environment

Publishing / Disseminating

- Version tracking, living papers
- FAIR (findable, accessible, interoperable, reusable) principles
- Metadata
 - Overall project, and all components
- Open Licenses
- Persistent identifiers (e.g. DOI, ORCID)

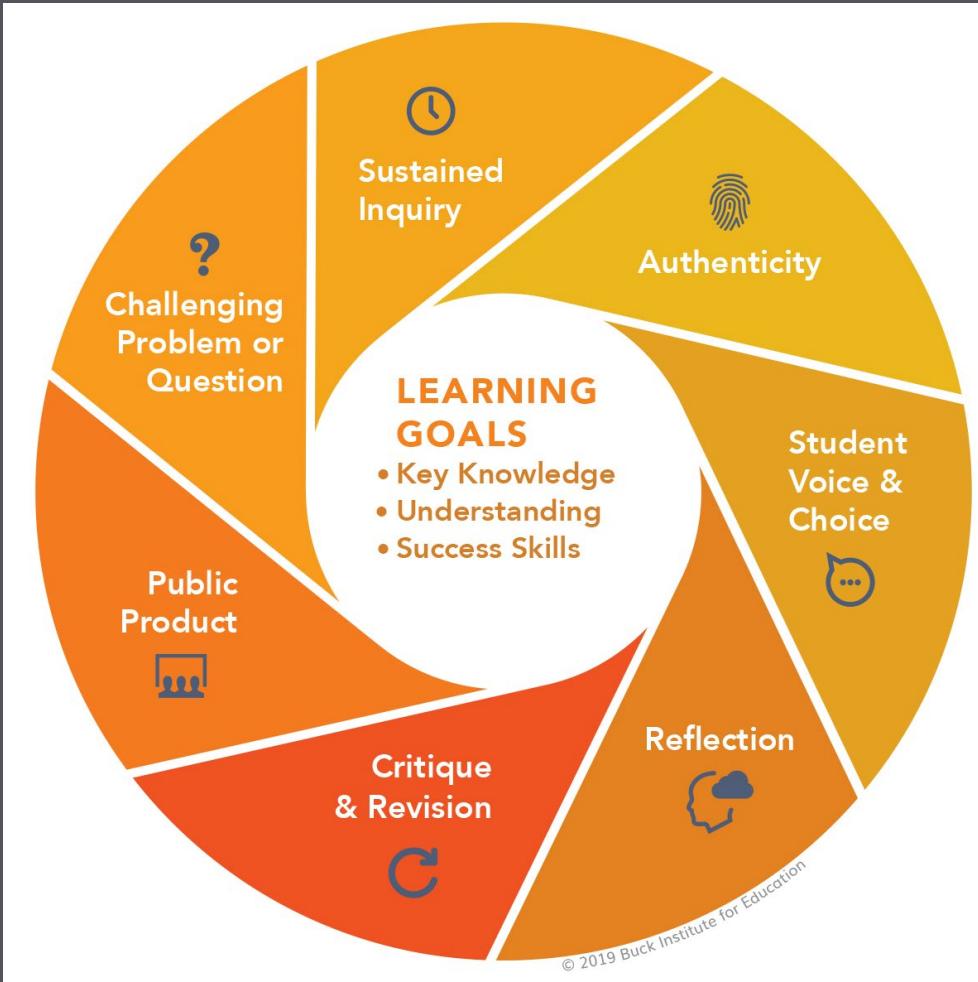
Infrastructure

- Curricula / pedagogy
- Git, GitHub, Markdown & Jekyll
 - Template Repository
- OSF / Figshare
- Python Jupyter Notebooks / R Markdown
- Docker, o2r.info
- Cyberinfrastructure, e.g. CyberGISX
- Data repository with access based on ethics review
- R&R publishing guidelines / article types / exemplar studies



R&R Pedagogy

Project-Based Learning



- Reproduction studies as project-based learning
 - Refocus learning goals on competencies
 - Improved engagement
 - Retain diversity in STEM
- Share motivations with students!
- PBL Gold Standard:
<https://www.pblworks.org/what-is-pbl/gold-standard-project-design>

Rethinking Learning Goals

- Software-specific methods & extensions? No...
- Typical GIS Project: theme, geodatabase & some maps
- Learn and practice competencies for:
 - Reading published research
 - Research planning / design
 - Learning new (open source) software
 - Comparing results / assessing error and uncertainty
 - Reporting findings

Open Source GIScience @ GIS4DEV.github.io

Open Source GIScience

Joseph Holler's Open
Source GIScience
Resources at
Middlebury College



GIS4DEV



HEGSRR



HEGSRR

Learning Goals

- Survey FOSS4G (Free and Open Source for Geospatial) in terms of its landscape of organizations and projects, research applications, and (radically) unique political economy of knowledge production.
- Expand your functional knowledge of the nature of geographic information with respect to data standards, structures, metadata, provenance, error, and uncertainty.
- Creatively apply Open GIScience to address compelling questions in human geography and problems in social and environmental sustainability.
- Critically reflect on emerging opportunities and ethical dilemmas in open-source geographic information science.
- Learn how to reproduce existing geographic research and to produce geographic research that is open, reproducible and replicable.
- Design and communicate research effectively in multiple media, including digital media, reports, presentations, maps, graphs, tables, data, and code.
- Become competent and confident in conducting research, learning new methods, and overcoming errors, uncertainty, and technical difficulties. Learn to "debug" problems and teach yourself new techniques through structured experimentation.

Discuss



- Reproducibility takes work to achieve...
(as if publishing research is not work enough)
 1. Why should we try? What are our individual and collective motivations?
 2. Any failures, concerns, misgivings?

Vision for R&R in Geography

- Studies are **preregistered**, conducted with **open source software**, published with permissible licenses, and attached to executable compendia
- Reproduction and replication studies are encouraged and published by preeminent journals
- Students learn contemporary methods and classic theories through reproductions and replications
- Students conduct and publish reproductions and replications, developing their methods while contributing to geographic sciences.

Break

- Have you created a GitHub account?
- Have you installed GitHub Desktop?



Part II: Research Compendium in Git & GitHub

GitHub Overview

Template for R&R Research in HEGS

GitHub Overview

Challenges

- Many versions of a research project over time
- Complex project management, especially with alternative possible solutions
- Collaborating with peers, research assistants
- Tracking contributions to the research
- Publishing / disseminating research compendium



```
528 559 race_gee <- geeglm(  
529     - covid_rate ~ white_pct + black_pct + native_pct + asian_pct + other_pct,  
530     + covid_rate ~ z_white_pct + z_black_pct + z_native_pct + z_asian_pct + z_other_pct,  
531     data = gee_data, # data frame  
532     id = id, # cluster IDs  
533     family = Gamma(link = "log"),  
534     @@ -537,7 +568,7 @@ race_gee <- geeglm(  
537 568 # coef() extracts coefficients table from the summary, same as $coefficients
```

- ... GitHub is a software company implementing open source Git

Git & GitHub Version Control

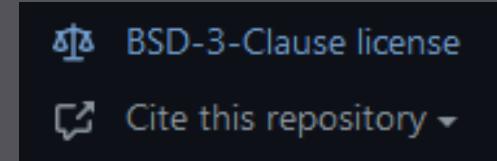


Command	Purpose
Clone	Download repo from GitHub to your computer
Fork	Copy repo into your remote GitHub account
Commit	Make a set of changes to a repo
Revert	Undo a commit
Fetch	Download remote commits from GitHub
Pull	Download & apply remote commits from GitHub
Push	Upload local commits to GitHub
Branch	Create a new distinct set of commits
Pull Request	Suggest, review & merge a fork or branch

GitHub Features



- Publishing: Markdown pages, GitHub pages with Jekyll
- Credit: LICENSE file and .CFF citation file
- Collaboration: Issues, Discussions, Wikis, Pull Requests
- Control & Organize Access: Public/Private, Organizations, Teams
- Version Releases: 1.0 Preanalysis, 1.1 Preprint, 1.2 Review...
- Difference visualization
- Free Services for GitHub Educators / Campus Advisors
- Integrations: Overleaf, OSF, RStudio, Atom ...



Tips, Tricks & Limitations



- Start Git repository with collaborators at project conception
- Commit frequently with good descriptions
- Use plain-text data formats
 - Markdown LaTeX rather than Microsoft Word
 - XML, CSV, JSON rather than Shapefile
 - One sentence on each new line
- No files > 100mb
- Never hard code passwords or API keys into your code
- Clear Jupyter notebook contents before committing changes

Template for R&R Research in HEGS

(Executable) Research Compendium

- Computational Notebook: Narrative and Code
- Data
- Processing Environment
- Metadata
- License
- Code executes/compiles into publication

Repository template for HEGS research

- [github.com/HEGSRR/
HEGSRR-Template](https://github.com/HEGSRR/HEGSRR-Template)
- Standard folder and metadata structure
- Template study pre-registration (original, reproduction, or replication)
- Reporting guidance that links preregistration with results
- Rmarkdown & Jupyter templates ...
- Integral to our pedagogy

 josephholler	drafted GEE code	...	5 hours ago	⌚ 57
 data	updates	yesterday		
 docs	Merge branch 'main' of https://github.com/HEGS...	15 days ago		
 procedure	drafted GEE code	5 hours ago		
 results	Initial commit	last month		
 .gitignore	Initial commit	last month		
 LICENSE	Initial commit	last month		
 r_project.rproj	Initial commit	last month		
 readme.md	update abstract in readme.md	last month		
 template_readme.md	Initial commit	last month		

Use .gitignore and code to manage data

DATA/	PRIVATE/	PUBLIC/
RAW/	ignore	
DERIVED/	ignore	
METADATA/		
SCRATCH/	ignore	

.gitignore example:

```
data/raw/private/**  
!readme.md
```

script example:

```
G =  
ox.graph_from_place('Illino  
is', network_type='drive',  
buffer_dist=24140.2)  
  
ox.save_graphml(G,  
'raw/private/  
Illinois_Network.graphml')
```

Practice: Managing GitHub Repositories

Please see the Workshop Technical Guide and follow along

Discuss



- Adopting a new technology to manage your scholarship may seem daunting...
 1. Are there any specific ways you want to integrate Git version control into your scholarship?
 2. Are there specific reasons that you wish you had already adopted Git in the past?
 3. Any concerns or misgivings?

Break

- Have you seen the original paper for our example reproduction 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):101007.
doi.org/10.1016/j.dhjo.2020.101007



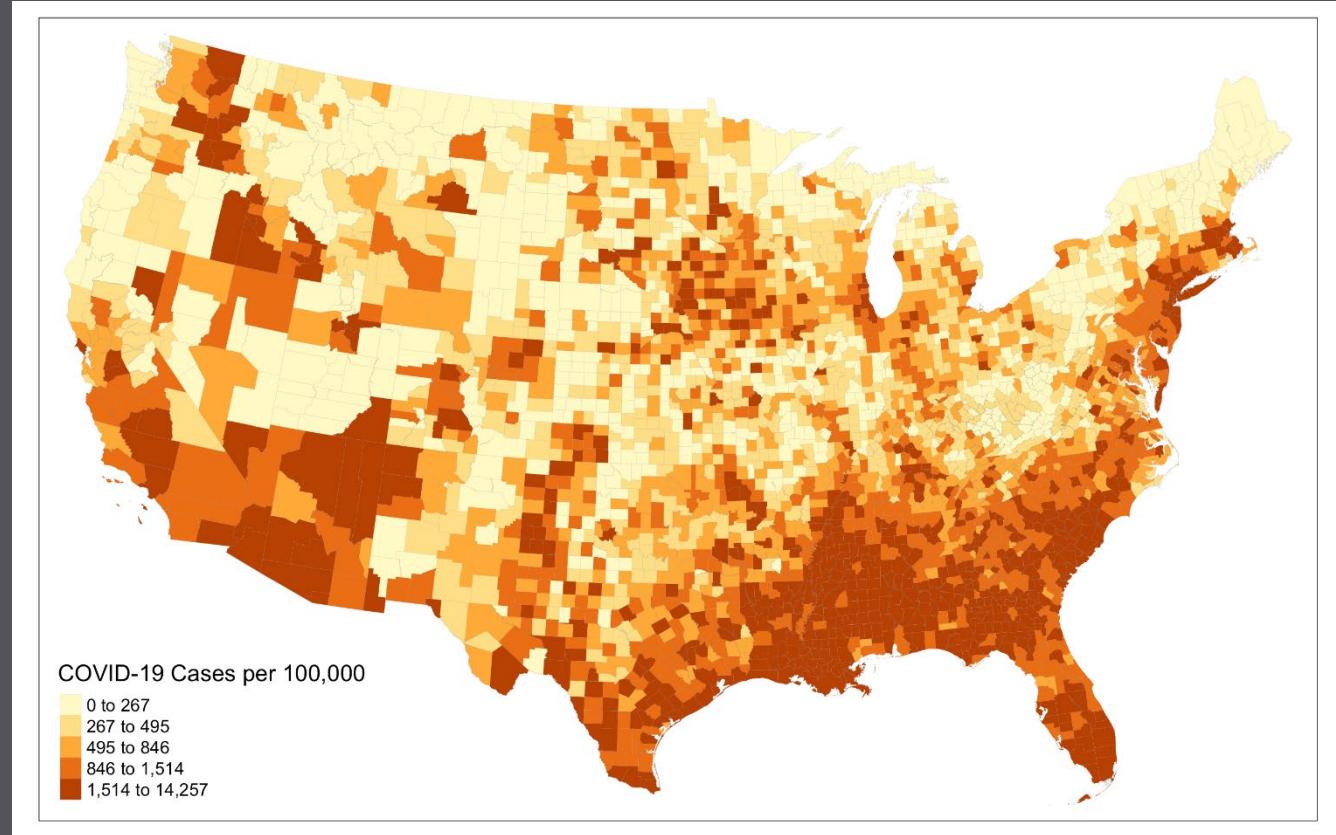
Part III: Plan and Register a Reproduction / Replication Study

Preanalysis Plan Overview

Pre-analysis Plan Overview

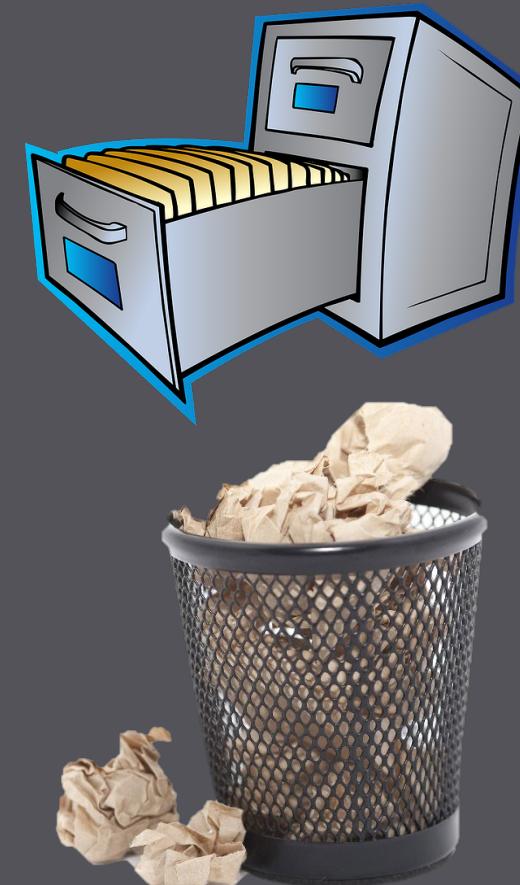
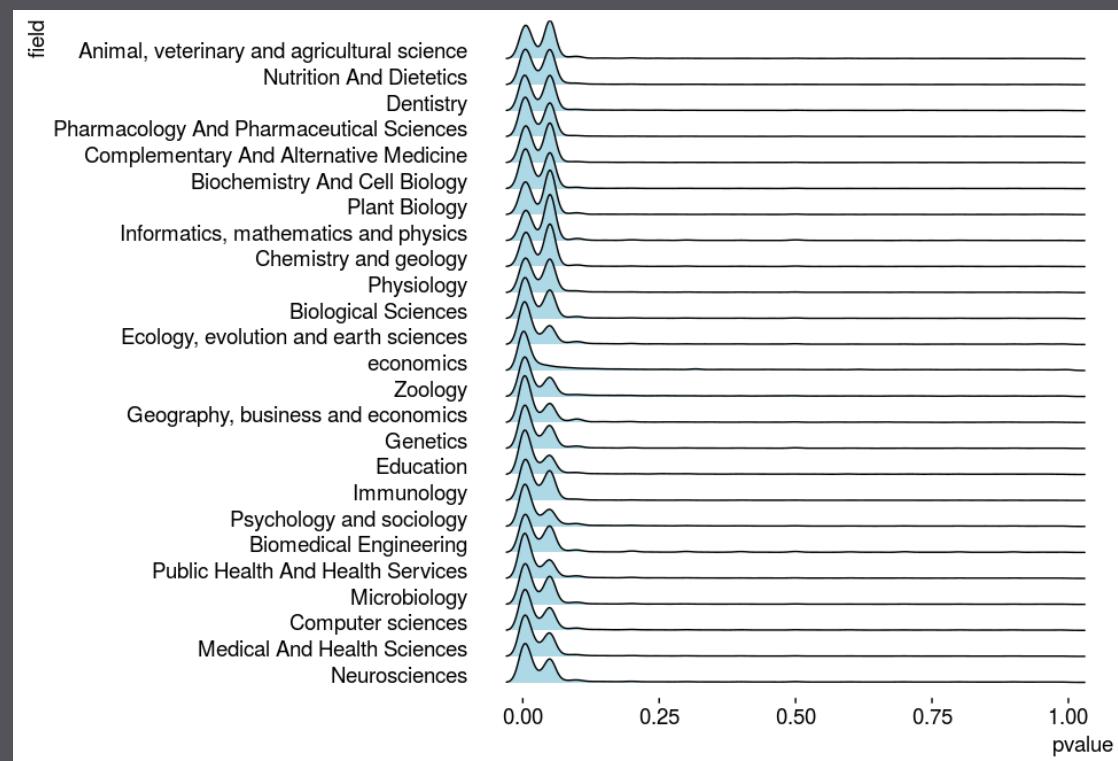
Reproduction study worked example

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):101007.



Pre-analysis Planning & Registration

- Challenge: P-hacking, Research File Drawer



OSF Pre-analysis Plan Registration

- <https://doi.org/10.17605/OSF.IO/S5MTQ>

The screenshot shows the OSF (Open Science Framework) interface. At the top, there is a navigation bar with the OSF logo, 'OSF HOME' with a dropdown arrow, 'My Projects', and a search bar. Below the navigation bar, there is a secondary navigation bar with tabs: 'Reproduction of Chakraborty 2021 Dist...', 'Files', 'Wiki', 'Analytics', 'Registrations' (which is highlighted in blue), 'Contributors', 'Add-ons', and 'Settings'. Under the 'Registrations' tab, there are two buttons: 'Registrations' (which is underlined in blue) and 'Draft Registrations'. On the right side of this row is a green button labeled 'New registration'. The main content area displays a registration card for a study. The title of the registration is 'Reproduction of Chakraborty 2021 Distribution of COVID-19 and intra-categorical analysis of people with disabilities'. Below the title, there are several metadata fields: 'Registration template: Open-Ended Registration', 'Registry: OSF Registries', 'Registered: Thu Jun 02 2022 15:05:02 GMT-0400', 'Last updated: Thu Jun 02 2022 15:02:52 GMT-0400', 'Contributors: Holler, Kedron, An-Pham, and 2 more', and 'Description: This study is a reproduction of Chakraborty, J. 2021. Social inequities in the distribution of COVID-19: An intra-categoric...'. Below the description, there is a 'Tags:' section with four tags: 'COVID-19', 'Disability', 'Reproduction Study', and 'United States'. At the bottom of the registration card are two buttons: 'View' and 'Update'. At the very bottom of the page, there is a footer message: 'To register the entire project "Reproducibility, Replicability, and Open Science Practices in the Geographical Sciences" instead, click [here](#)'.

OSF Pre-analysis Plan Registration

- Log in with ORCID
- Create a project (public or private)
- Link to GitHub repository (Add-ons → GitHub)
<https://help.osf.io/article/211-connect-github-to-a-project>
- Go to Registrations → New Registration
 - Title, Description, Contributors, Category (procedure, project), License (BSD 3-Clause)
- Will be archived & appear under “registrations”
- Can be embargoed, but not retracted or changed.
- Generate DOI for any public OSF project or project component

R&R Pre-Analysis Plan

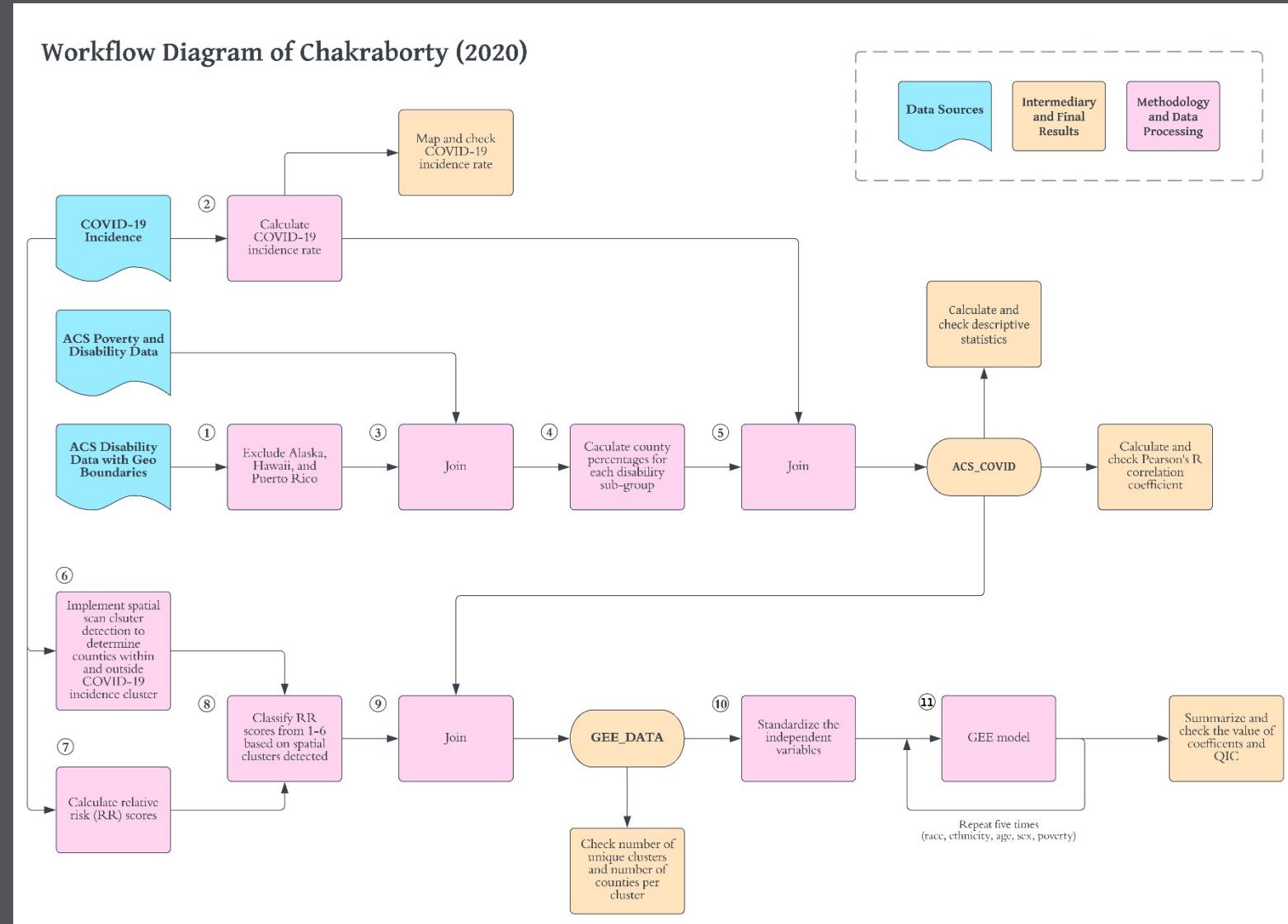
- Original Study Information
 - Abstract
 - Geographic extent
 - Spatial Support
 - Data sources & their availability
- Analytical Plan
 - Original Data Sampling
 - Secondary Data
 - Variables, including attribute or geographic transformations
 - Analytical specification (e.g. statistical model parameters)
 - Inference criteria, results & robustness
- Protocol & any planned differences from original study

Pre-analysis Planning

- Close reading of paper, highlighting
 - Data Sources / Inputs
 - Methodology and Data Processing / Methods / Tools
 - Intermediary and Final Results / Outputs
- Research data sources & metadata *without viewing data directly*
- Research / review unfamiliar methods, especially assumptions, inputs, parameters, outputs

Workflow Diagram

- Circular data layers
 - Rectangular processes
 - Interactive activity /w notecards & cutout figures



Activity

- Gather in small groups
- Number each paragraph of the research paper
- Annotate the workflow diagram, indicating which paragraphs of the paper contain information about that analytical step or data source
- This is why computational notebooks are valuable for reproducibility— interweave narrative and code!

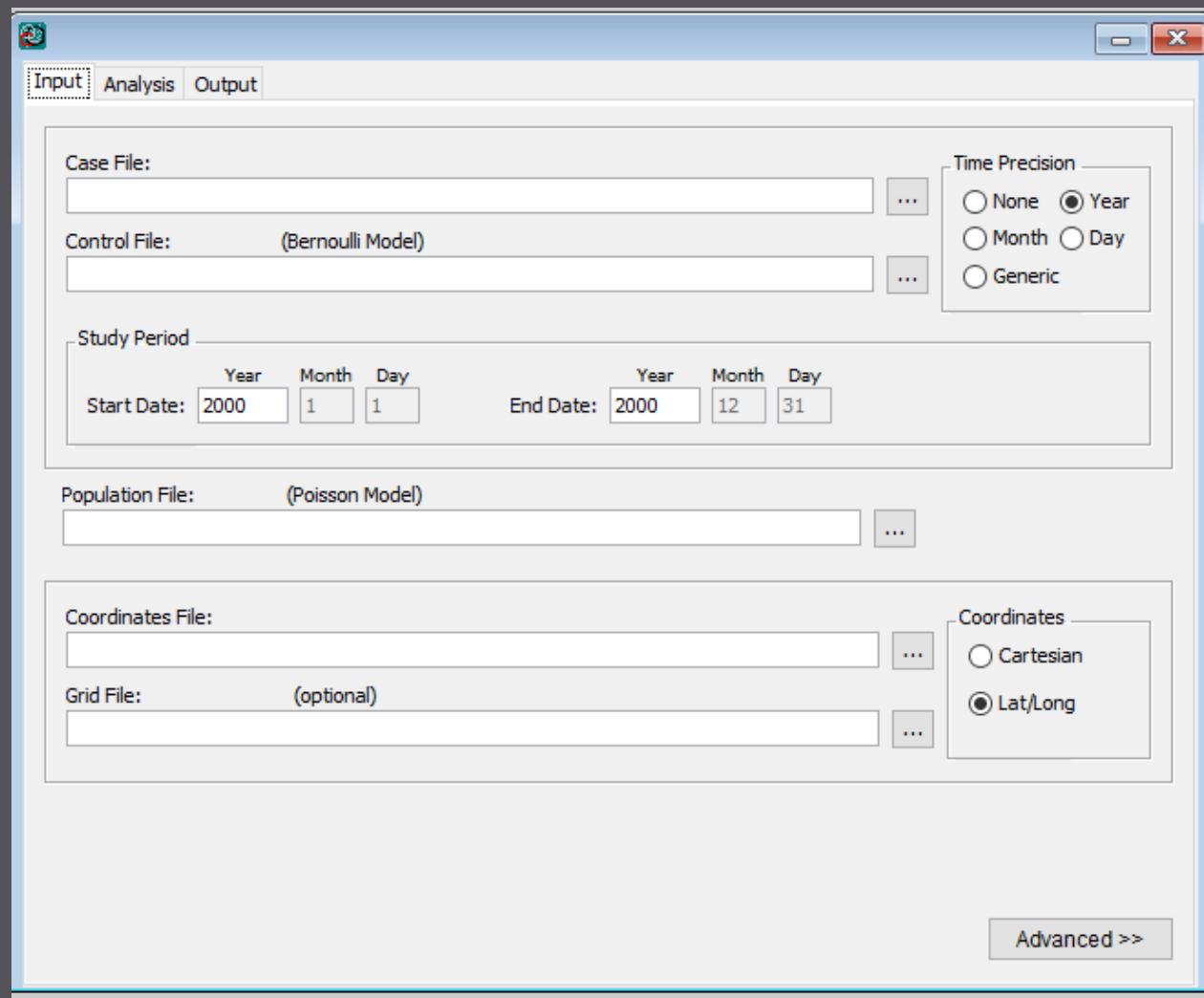
The Spatial Scan Statistic: SaTScan & SpatialEpi

SaTScan

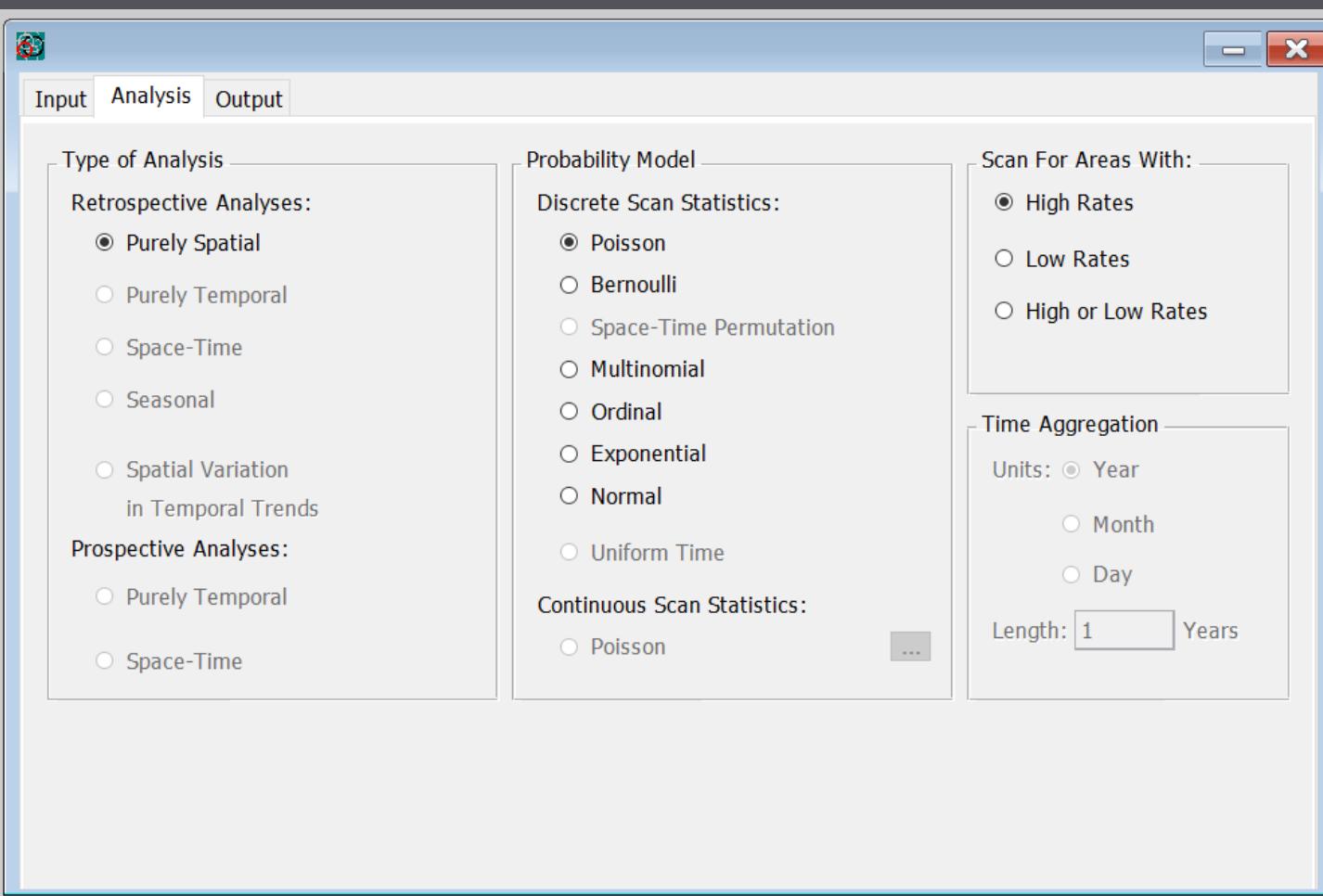
- Free
- Not open source
- Requires three input files and many parameters
- <https://www.satscan.org>
- Reproducible?
 - Requires manual use of multiple programs or...
 - Requires separate installation of SaTScan and system-specific commands from within R
- SaTScan™ is a trademark of Martin Kulldorff. The SaTScan™ software was developed under the joint auspices of (i) Martin Kulldorff, (ii) the National Cancer Institute, and (iii) the New York City Department of Health and Mental Hygiene.



SaTScan Input



SaTScan Analysis



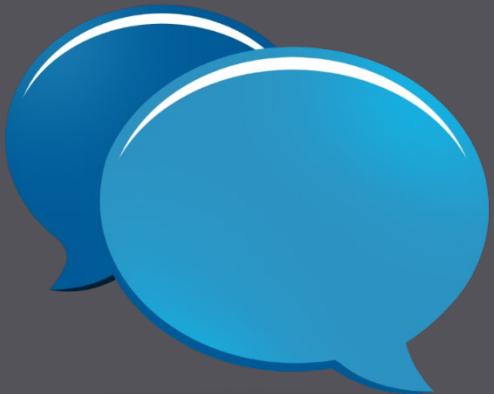
SaTScan Advanced Options

- Check data for temporal extent
- Check data for geographic extent
- Define non-Euclidean spatial neighbors
- Use network space

Generalized Estimating Equations

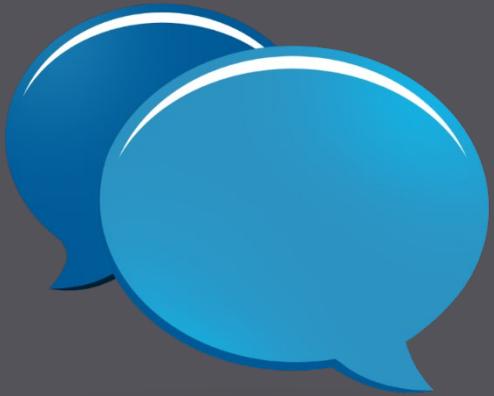
- Not inferential: approximate relationships
- Extension of GLM commonly applied to longitudinal panel data to account for dependency in a panel subject's data over time
- Assumes *correlation* within clusters and *independence* between clusters
- Model Specification
 - Intra-cluster dependency matrix
 - Distribution
 - Link function

Discuss



- This paper is very precise and clear... but...
 1. Have we already learned more from researching its data sources and planning a workflow? What?
 2. Do you notice any errors or missing information? If you were peer reviewing this study or planning to reproduce it, what more would you want to know?
 3. Have you developed any concerns about the study?

Discuss



- Pre-analysis plans and workflows are intended to increase transparency and keep researchers honest...
- 1. Does the “spatial is special” phrase apply to pre-analysis plan registration?
- 2. Are there ways in which spatial data and spatial analysis require additional attention to detail and scrutiny if pre-analysis planning is to support enhanced replicability?

Break

- Do you have a Census API key?
- Have you already installed the latest versions of R and RStudio?



Part IV: Execute a Computational Notebook in RStudio

RStudio & R Markdown Overview

Rstudio: GUI for R

The screenshot displays the RStudio graphical user interface. The main window shows a R Markdown file titled "01-RPr-Chakraborty.Rmd" with code and text. A red box highlights the code area, and an orange box highlights the "coefficient_results" tab in the environment panel. The environment panel lists various R objects with their details. A black box highlights the "Currently displaying files" section. The R console at the bottom shows the R version and a command to view a dataset.

Main Window

Environment Panel

R Console

github.com/HEGSRR OSF.IO/C5A2R

506:79 C Chunk 23: preprocess data for GEE model

R Pr-Chakraborty-2021 - main - RStudio

Intermediaries from environment panel

gpk gpk

48 The data on disability and sociodemographic characteristics come from the U.S. Census American Community Survey (ACS) five-year estimates for 2018 (2014–2018).
49
50 There is no **randomization** in the original study.
51
52 The study was originally conducted using SatScan software (unspecified version) to implement the spatial scan statistic.
53 Other software are not specified in the publication; however data files and communication with the author show that spatial analysis and mapping was conducted in ArcGIS and statistics were calculated in SPSS.
54
55
56
57 ````{r setup, message = FALSE, include = FALSE}`
58 # list of required packages
59 packages = c("tidycensus", "tidyverse", "downloader", "sf", "classInt", "readr",
60 "here", "s2", "pastecs", "tmap", "SpatialEpi", "svDialogs",
61 "geepack")
62
63 # load and install required packages
64 package.check <- lapply(
65 packages,
66 FUN = function(x) {
67 if (!require(x, character.only = TRUE)) {
68 install.packages(x, dependencies = TRUE, quietly=TRUE)
69 library(x, character.only = TRUE)
70 }
71 }
72)
73

506:79 C Chunk 23: preprocess data for GEE model

Console Terminal Jobs

R 4.1.1 · ~/Documents/GitHub/HEGSRR/RPr-Chakraborty-2021/ ↵

+
>
> View(covid_kulldorff)

Environment History Connections Git Tutorial

Import Dataset 37 MB List C

cluster_table 96 obs. of 16 variables
coefficient_re... 22 obs. of 4 variables
covid 3108 obs. of 25 variables
covid_kulldorff Large list (5 elements, 47.3 MB)
covid_rate_tab... 3108 obs. of 2 variables
covid_table 3108 obs. of 6 variables
covid_temp 3108 obs. of 26 variables
ethnicity_gee List of 28
gee_data 3059 obs. of 45 variables
i List of 8
original_clust... 96 obs. of 17 variables
package.check List of 13

Files Plots Packages Help Viewer More options to display plots, packages...

New Folder Delete Rename More

Name Size Modified

.. 23.5 KB Jun 1, 2022, 2:09 PM
01-RPr-Chakraborty.Rmd 141 B May 31, 2022, 1:04
readme.md 7.3 KB May 31, 2022, 1:04
Scratch-Code.Rmd

Currently displaying files

74

R Markdown Computational Notebook

The screenshot shows an RStudio interface with the following code in an R Markdown file:

```
13
74 # save the R processing environment
75 writeLines(capture.output(sessionInfo()),
76             here("procedure", "environment", "r_environment.txt"))
77 ``
78
79 ## Query American Community Survey Data
80
81 This will require an API key for the census, which can be acquired easily here:
[api.census.gov/data/key_signup.html](https://api.census.gov/data/key_signup.html)
82 This query can take some time to run...
83
84 ```{r Load ACS Data, message = FALSE, eval = FALSE}
85 # get API Key
86 # we could store this in the raw/private or scratch folder and load if the
87 # researcher has already entered it once
88 census_api_key(dlgInput("Enter a Census API Key",
89   Sys.getenv("CENSUS_API_KEY"))$res,
90   overwrite = TRUE)
91
92 # Query disability demographic data with geographic boundaries
93 acs <- get_acs(geography = "county",
94   table = "S1810",
95   year = 2018,
96   output = "wide",
97   cache_table = TRUE,
98   geometry = TRUE,
99   keep_geo_vars = TRUE)
100
```

The code is annotated with three boxes:

- Text**: Lines 79-82, highlighted with a blue border.
- Code block**: Lines 84-90, highlighted with a red border.
- Comments**: Line 92, highlighted with an orange border.

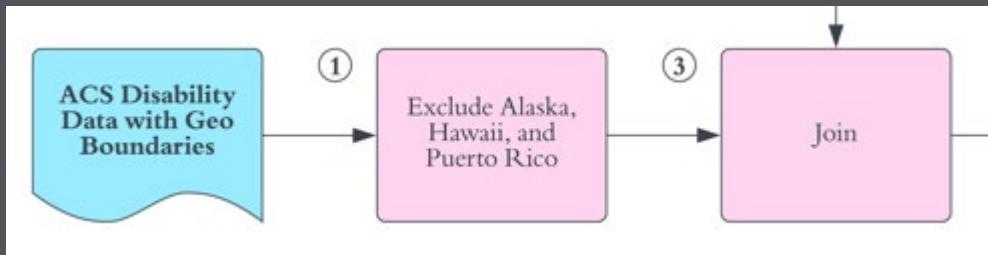
Annotations in the RStudio interface:

- "Add a code block here" and "Alternatively, select the lines and run the code here" are located in the toolbar.
- A "Run the code here" button is located in the sidebar.

Bottom status bar: 506:79 | Chunk 23: preprocess data for GEE model | R Markdown

Reading R

- **filter and join acs data code block**
- # Remove Alaska, Hawaii & Puerto Rico
- ```
acs <- filter(acs, !STATEFP %in%
c("02", "15", "72"))
```
- # Join poverty data to disability data
- ```
acs <- left_join(acs, acs_pov, by =  
"GEOID")
```



- **Alternative with %>% piping**
- ```
acs <- acs %>%
filter(!STATEFP %in% c("02", "15",
"72")) %>%
left_join(acs_pov, by = "GEOID")
```

# R / R Markdown Practices

- open project in root directory of compendium to begin
- load and install packages automatically (see our setup code block)
- record the processing environment (see our setup code block)
- use the **here** package for relative path names
- write code in Rmarkdown or R scripts, not the console!
- save API keys, passwords in local environment or scratch directory that is not version-tracked
- use discrete code chunks that view & check results
- Start with code to download data directly or through API
- Save data so that code can run without downloading
- save important inputs and results as individual objects in .rds files:  
Do not rely on the .RData file or objects stored in the environment
- save figures with code, not manually
- **tidyverse** style guide & helper functions for legible code
- **Knit** Rmarkdown into PDF with LaTeX

# Practice: Execute R Markdown Computational Notebook

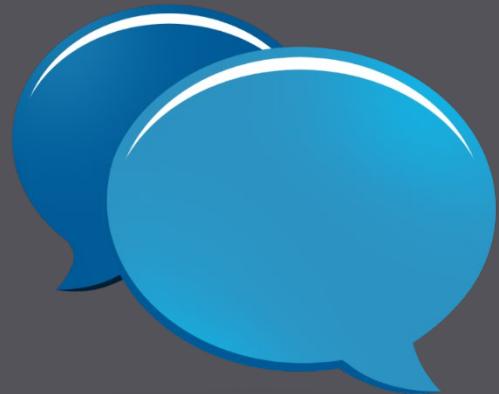
Please follow along with the Workshop Technical Guide

# Reproduction Study Results

# Comparing GEE Results

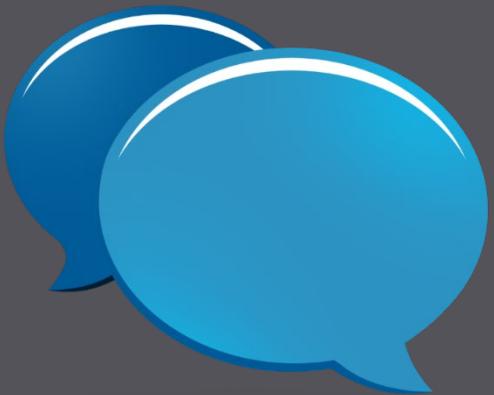
|                        | Our Coef | Orig Coef | Coef Diff | Our sig | Orig sig* | Our QIC | Jay QIC | QIC Diff |
|------------------------|----------|-----------|-----------|---------|-----------|---------|---------|----------|
| <b>race</b>            |          |           |           |         |           |         |         |          |
| (Intercept)            | 7.72     | 7.11      | 0.62      | < 0.001 | < 0.01    | 2616.4  | 2582.5  | 33.9     |
| white_pct              | -0.13    | -0.20     | -0.07     | < 0.001 | < 0.01    |         |         |          |
| black_pct              | 0.02     | 0.11      | -0.09     | < 0.05  | < 0.01    |         |         |          |
| native_pct             | 0.02     | 0.05      | -0.03     | < 0.001 | < 0.01    |         |         |          |
| asian_pct              | 0.02     | 0.08      | -0.06     | < 0.001 | < 0.01    |         |         |          |
| other_pct              | 0.02     | 0.08      | -0.06     | < 0.001 | < 0.01    |         |         |          |
| <b>ethnicity</b>       |          |           |           |         |           |         |         |          |
| (Intercept)            | 7.72     | 7.19      | 0.53      | < 0.001 | < 0.01    | 2616.3  | 2586.6  | 29.8     |
| non_hisp_white_pct     | -0.15    | -0.24     | -0.09     | < 0.001 | < 0.01    |         |         |          |
| hisp_pct               | 0.01     | 0.12      | -0.11     | 0.198   | < 0.01    |         |         |          |
| non_hisp_non_white_pct | 0.02     | 0.12      | -0.10     | < 0.01  | < 0.01    |         |         |          |
| <b>poverty status</b>  |          |           |           |         |           |         |         |          |
| (Intercept)            | 7.77     | 7.18      | 0.59      | < 0.001 | < 0.01    | 2562.7  | 2801.5  | -238.8   |
| bpov_pct               | 0.02     | 0.15      | -0.13     | < 0.01  | < 0.01    |         |         |          |
| apov_pct               | -0.11    | -0.27     | -0.16     | < 0.001 | < 0.01    |         |         |          |
| <b>age</b>             |          |           |           |         |           |         |         |          |
| (Intercept)            | 7.78     | 7.24      | 0.54      |         | < 0.01    | 3577.1  | 2978.7  | 598.3    |
| pct_5_17               | 0.02     | 0.05      | -0.03     | < 0.001 | < 0.01    |         |         |          |
| pct_18_34              | 0.01     | 0.04      | -0.02     | < 0.001 |           |         |         |          |
| pct_35_64              | -0.02    | -0.03     | 0.00      | < 0.01  |           |         |         |          |
| pct_65_74              | -0.06    | -0.09     | -0.03     | < 0.001 | < 0.01    |         |         |          |
| pct_75                 | -0.05    | -0.11     | -0.06     | < 0.001 | < 0.01    |         |         |          |
| <b>biological sex</b>  |          |           |           |         |           |         |         |          |
| (Intercept)            | 7.78     | 7.22      | 0.56      | < 0.001 | < 0.01    | 2012.3  | 2892.4  | -880.1   |
| male_pct               | -0.14    | -0.30     | -0.16     | < 0.001 | < 0.01    |         |         |          |
| female_pct             | 0.04     | 0.15      | -0.11     | < 0.001 | < 0.01    |         |         |          |

# Reproduction Report



- Append three sections to the pre-registered analysis plan
  - Results
  - Unplanned Deviations
    - Did you noticed any deviations from the research plan in the R Markdown notebook?
  - Discussion
    - Was this reproduction an exact success? Approximate success? Failure?
    - Why?

# Discuss



- This paper is very precise and clear...  
but...
- 1. Have we learned anything about Chakraborty's study that *we did not know* based on the paper? What?
- 2. Can you imagine ways in which the reproduction study could be replicated, reanalyzed, or extended?

# Beyond Reproduction

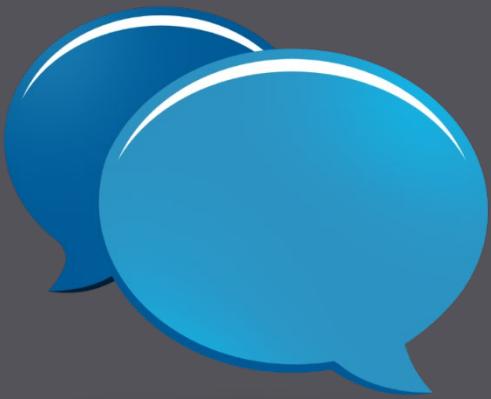
|                | Same Methods                   | Varied Methods |
|----------------|--------------------------------|----------------|
| Same Data      | Reproduction<br>(Verification) | Reanalysis     |
| Different Data | (Direct) Replication           | Extension      |

# Discuss



- Is there a future for R&R in human-environment and geographical sciences for you?
  1. Are there specific ways you can imagine benefitting from R&R practices into your ongoing and future scholarly activity?
  2. Are there specific barriers to adopting R&R practices into your ongoing and future scholarly activity?

# Discuss



1. If we can work out R&R in geography, can you envision ways in which geography could contribute to other sciences or to convergent / interdisciplinary research?
2. Interest in trying a reproduction or replication project in a course you are teaching?

# Appreciation

- National Science Foundation
- UCGIS
- Jayajit Chakraborty
- Peter Kedron
- Research Assistants & Independent Studies: Kufre Udo, Derrick Burt, Sarah Bardin, Drew An-Pham, Emily Zhou, Maddie Tango
- We welcome feedback on all materials shared today! Email, Github Issue, or Pull Request!